



Thèse

**présentée pour obtenir le grade de docteur
de l'Ecole Nationale Supérieure
des Télécommunications**

Spécialité : Informatique et Réseaux

François Yvon

**Prononcer par analogie : motivation,
formalisation et évaluation**

ENST 96 E 015

Soutenue le 14 mai 1996 devant le jury composé de

Jean-Paul Haton

Président

Eric Laporte

Rapporteurs

Laurent Miclet

Alain Bonnet

Examineurs

Gérard Chollet

Anne Lacheret-Dujour

Ecole Nationale Supérieure des Télécommunications
Enseignement supérieur des Télécommunications



Thèse

**présentée pour obtenir le grade de docteur
de l'Ecole Nationale Supérieure
des Télécommunications**

Spécialité : Informatique et Réseaux

François Yvon

**Prononcer par analogie : motivation,
formalisation et évaluation**

ENST 96 E 015

Soutenue le 14 mai 1996 devant le jury composé de

Jean-Paul Haton

Président

Eric Laporte

Rapporteurs

Laurent Miclet

Alain Bonnet

Examineurs

Gérard Chollet

Anne Lacheret-Dujour

Ecole Nationale Supérieure des Télécommunications
Enseignement supérieur des Télécommunications

Remerciements

Je remercie monsieur le professeur Jean-Paul Haton, de l'Université de Nancy, de m'avoir fait l'honneur de présider, avec délicatesse et décontraction, mon jury de thèse. Je remercie également messieurs les professeurs Éric Laporte, de l'Université de Reims, et Laurent Miclet, de l'ENSSAT d'avoir accepté de tenir, chacun dans son registre, le délicat rôle de rapporteur : leurs critiques et remarques m'ont été précieuses. Je remercie Madame Anne Lacheret-Dujour, maître de conférence à l'Université de Caen, pour l'intérêt sincère dont elle a témoigné vis-à-vis de mes travaux. Je remercie Gérard Chollet, directeur de recherche CNRS à l'ENST, pour son soutien et son aide dans les derniers moments de cette thèse : ils ont été les bienvenus. Merci enfin au professeur Alain Bonnet, de l'ENST, d'avoir bien voulu, en dépit de l'éloignement, continuer à superviser ce travail.

Martin Rajman, avec qui j'ai eu l'avantage de cohabiter (à mi-temps!) pendant ces trois et quelques années, et dont j'ai pu apprécier au quotidien la gentillesse, la compétence et la disponibilité, mérite toute ma reconnaissance : il l'a sans réserve. Merci également à Ulrich Finger et à tous les membres du département informatique de l'ENST, qui ont contribué, chacun à sa manière, à ce que ce travail se déroule dans un environnement agréable, en particulier à Jean-Louis Dessalles, Alain Grumbach, Jean-Cédric Chappelier, Jean-Bernard Auriol, Arnon Rungsawan, Denis Bonnet, Jacques Han, et Didier Verna, avec qui les interactions scientifiques ont souvent été stimulantes, mais aussi à Nathalie Leruyet, Olivier Hudry, Antoine Lobstein, Frédéric Woïrgard, Irène Charon, Anne Germa, Philippe Picouet, Laurent Pautet, Vincent Oria, Frédéric Château, Francette Delbor, et j'en oublie sûrement, pour leur amical soutien.

Durant ces quelques années de travail, j'ai eu la chance de développer des relations de travail cordiales et riches avec Frédéric Bimbot et Sabine Deligne, du département Signal de l'ENST, avec Vito Pirrelli et Stefano Federici de l'Istituto di Linguistica Computazionale de Pise, avec Valérie Beaudoin de l'EHESS, ainsi qu'avec tous les partenaires du projet Européen ONOMASTICA : à tous et à toutes j'exprime ma profonde et sincère gratitude.

Puisque de noms propres il va être amplement question, qu'il me soit permis d'ajouter à cette déjà longue liste ceux de Thierry Lambert, d'Éric Davalo, de Guillaume Deffuant, d'Emmanuel Guerre, de Christophe Tollu, et de Saadi Lahlou; année après année, leur soutien chaleureux et stimulant ne m'a jamais fait défaut.

Sommaire

Notations	11
Introduction	15
Motivations et objectifs	15
Organisation générale	18
1 Présentation du contexte	21
1.1 La transcription graphème-phonème	21
1.2 La plausibilité cognitive de l'hypothèse analogique pour la lecture	39
1.3 Quelques notes sur l'analogie	45
1.4 Conclusion	49
2 Prononcer/Classer	51
2.1 Préambule	51
2.2 Apprentissage de la transcription graphème-phonème	54
2.3 Transcription à base de lexique	67
2.4 À propos des approches classificatoires	74
3 Prononcer par recollement de segments	85
3.1 Préambule	85
3.2 Transcription par recombinaison binaire de segments	88
3.3 Transcription par recombinaison n-aire de segments	104
3.4 Le modèle multigramme à deux niveaux	119
3.5 Conclusion sur les approches par recollement de segments	133
4 Analogies et paradigmes	137
4.1 Préambule	137

4.2 Morphologie, Analogie, Phonologie	141
4.3 Prononcer par construction d'analogies paradigmatiques	151
4.4 Prononcer par application de transductions paradigmatiques . . .	161
4.5 La conjugaison des verbes anglais	193
4.6 Conclusion	205
Conclusion	209
Annexes	213
A Les mécanismes d'alignement	213
A.1 Position du problème	213
A.2 Quelques méthodes d'alignement	214
A.3 Quelques expériences d'alignement du lexique NETTALK	220
B DEC : un système classificatoire prototypique	223
B.1 Un exemple introductif	223
B.2 Examen de l'algorithme	226
B.3 Quelques résultats	229
B.4 Perspectives et conclusions	234
C Le système PRONOUNCE	237
C.1 Construction du graphe d'analogies	237
C.2 Classement des solutions	238
C.3 Les performances réelles de PRONOUNCE	240
C.4 Conclusion	241
D Protocole général d'évaluation des performances	243
D.1 Mesure des performances	243
D.2 Comparaison de deux algorithmes	244
E Le lexique NETTALK	247
F Le lexique BDLEX	251
G Les lexiques d'ONOMASTICA	255
G.1 Le lexique ENST	255

Sommaire	5
G.2 Le lexique ILC	256
G.3 Le lexique KUN	256
G.4 Le lexique TUB	256
H Le lexique CMU	259
I Le lexique de verbes anglais	263
Bibliographie	266

Table des figures

1.1	Un modèle de double accès lexical.	41
1.2	L'analogie comme un rapport de proportion	47
2.1	Quelques alignements de <i>chanson</i> avec /fãs̃/.	52
2.2	Architecture de Nettek	56
2.3	Architecture de Dietterich et Bakiri.	60
2.4	Taux de réussite de DEC en fonction du score.	76
2.5	Évaluation de la vitesse d'apprentissage de DEC sur les lexiques ENST et CMU.	82
3.1	Un mot plus un mot égale un mot.	89
3.2	Taux de correction en fonction de la taille du recouvrement.	93
3.3	Courbe d'apprentissage sur le lexique ENST.	97
3.4	Un exemple de recombinaison erronée.	98
3.5	Extraction des sous-chaînes du mot <i>abash</i>	106
3.6	Construction du graphe d'analogies	108
3.7	Critère principal de classement des solutions de SMPA	110
3.8	Courbes d'apprentissage sur les lexiques CMU et ENST	115
3.9	Évolution des taux de correction et de silence en fonction de la contrainte de recouvrement.	116
3.10	Performances sur la tâche de transcription en fonction des contraintes de taille	127
4.1	Un système d'alternances	139
4.2	Un réseau d'alternances	139
4.3	Extension paradigmatique	140
4.4	La segmentation associée à une construction analogique.	155
4.5	Évaluation de la vitesse d'apprentissage de PPP sur le lexique ENST.	158

4.6	Un graphe de commutation morpho-syntaxiquement contraint. .	159
4.7	Un double graphe de commutation.	164
4.8	Un graphe de commutation tri-dimensionnel.	166
4.9	Un graphe de commutation abstrait.	167
4.10	L'arbre des dérivés du mot cible <i>lecteur</i> , avant élagage.	176
4.11	L'arbre des dérivés du mot cible <i>lecteur</i> , après élagage.	179
4.12	Évaluation de la vitesse d'apprentissage de PCP sur le lexique ENST.	187
4.13	Pourcentage de succès en fonction du nombre d'analogues.	189
A.1	Différents alignements de <i>voyager</i> avec /vwajaze/.	214
A.2	Modèle de phonème pour l'alignement	217
A.3	Le critère d'alignement en fonction du nombre d'erreurs.	221
B.1	Évaluation de la vitesse d'apprentissage de DEC sur le lexique NETTALK.	232
B.2	Évaluation de la vitesse d'apprentissage de DEC sur le lexique ENST.	233
C.1	PRONOUNCE : construction du graphe d'analogies.	238

Liste des tableaux

0.1	Symboles phonétiques de l'alphabet phonétique international . .	13
2.1	Évaluation d'une stratégie aveugle au contexte.	78
3.1	Les segments extraits du mot <i>chanson</i>	88
3.2	Matrice de confusion entre les transcriptions de DEC et JUPA. . .	94
3.3	Quelques mots imprononçables.	95
3.4	Performances de JUPA lorsque le lexique croît.	97
3.5	Performances de JUPA sur différents lexiques.	103
3.6	Performances de SMPA en fonction de la taille des segments. . . .	112
3.7	Performances de SMPA sur différents lexiques.	113
3.8	Matrice de confusion entre les transcriptions de DEC et SMPA. . .	114
3.9	Matrice de confusion entre les transcriptions de JUPA et SMPA. .	114
3.10	Quelques cosegmentations de la lexie <i>trois</i>	122
3.11	Performances en décodage pour différentes valeurs des contraintes de taille	127
3.12	Performances en décodage pour le modèle (5,3) pour différents seuils d'écrtage	128
3.13	Performances en décodage pour différentes contraintes de taille, avec initialisation.	129
3.14	Quelques transcriptions produites par application du modèle (4,3).	131
3.15	Quelques transcriptions produites par application du modèle (4,3) avec modèle auxiliaire.	132
4.1	Quelques paradigmes patronymiques du lexique français.	148
4.2	Performances de PPP sur différents lexiques.	157
4.3	Un exemple de transformation finale.	172
4.4	Quelques transformations problématiques	173

4.5	Un ensemble de transformations pour lire <i>lecteur</i>	175
4.6	Les probabilités des transformations permettant de lire <i>lecteur</i> . .	182
4.7	À la recherche de <i>section</i>	184
4.8	Performances de PCP sur différents lexiques.	186
4.9	Quelques mots sans analogue.	189
4.10	Quelques analogies extraites par PCP.	190
4.11	État des performances sur l'apprentissage de la conjugaison de l'anglais	198
4.12	Performances de PCP pour les trois tâches de conjugaison. . . .	199
4.13	Performances de PCP pour les trois tâches de conjugaison. . . .	202
A.1	Quelques correspondances graphémiques de la séquence /ai/ . . .	215
B.1	Apprentissage par DEC des règles de lecture de la lettre a	224
B.2	Règles extraites pour la prononciation des lettres a et n	226
B.3	Performances de DEC en fonction de différentes stratégies d'ex- pansion du contexte.	229
B.4	Valeurs moyennes de l'information mutuelle	230
B.5	Les erreurs les plus fréquemment commises par DEC.	231
B.6	Performances de DEC sur différents lexiques	234
C.1	Performances de PRONOUNCE en fonction de la taille des segments.	240
C.2	Performances de PRONOUNCE sur différents lexiques.	241
E.1	Alphabet phonétique utilisé dans le lexique NETTALK	249
F.1	Quelques exemples tirés de BDLEX	252
F.2	Quelques exemples tirés de BDLEX	252
F.3	Alphabet phonétique utilisé dans le lexique BDLEX	254
G.1	Alphabet phonétique utilisé dans le lexique ENST	257
H.1	Alphabet phonétique utilisé dans le lexique CMU (les accents ont été omis)	261
I.1	Répartition des formes verbales par flexion	263
I.2	Alphabet phonétique Unibet utilisé dans le corpus de verbes anglais	265

Notations

Tout au long de ce document, nous serons amené à utiliser les concepts de *graphèmes* et de *phonèmes*. Afin de lever les ambiguïtés possibles sur le sens exact que nous donnons à ces termes, les quelques précisions suivantes nous semblent nécessaires.

Le concept de *graphème* possède en linguistique deux acceptions distinctes, qui relèvent de deux conceptions relativement antagonistes du système d'écriture. Ces deux acceptions correspondent aux définitions suivantes, tirées de (Mounin 1974) : « A. Unité graphique minimale entrant dans la composition de tout système d'écriture, le graphème peut représenter un concept (écriture idéographique) ou un élément de la réalisation phonique (écriture phonographique : syllabique, consonantique, alphabétique). Dans l'écriture alphabétique, il est communément appelé lettre. B. Signe substitutif au phonème dans le système graphique de représentation d'une langue : le groupement des deux lettres "q" et "u" constitue en français le graphème "qu" représentant le phonème /k/. » Dans l'ensemble de ce document, nous avons utilisé ce terme dans le sens A précédemment défini, c'est-à-dire en faisant abstraction du rapport entre les unités graphiques et le système de codage approximatif de la prononciation qu'elles réalisent. Pour référer au second sens de *graphème*, nous utiliserons préférentiellement le terme de *phonogramme* (Catach 1984). Comme nous l'employons, *graphème* se distingue assez peu de *lettre*, ou *caractère alphabétique*, qui seront parfois utilisés comme des synonymes. Notons toutefois que la concordance n'est pas exacte, puisque des caractères non-alphabétiques peuvent appartenir à ce système, comme le point (utilisé en français dans les sigles et abréviations), ou le tiret. Quelques termes afférents : *graphémique*, *graphique*, *orthographique*, *graphie*.

Les symboles graphémiques rencontrés dans les lexiques que nous avons manipulé sont les suivants :

a, à, â, ä, e, é, è, ë, ê, î, ï, ï, o, ô, ö, œ, u, û, ü, ù, y, , b, c, ç, d, f, g, h, j, k, l, m, n, , p, q, r, s, ß, t, p, s, v, w, x, z, -, ', .

Le concept de *phonème* possède en linguistique le sens précis de « son d'une langue, défini par des propriétés distinctives (traits pertinents) qui l'opposent aux autres sons de cette langue » (Petit Larousse). C'est en conséquence un concept abstrait, qui ne peut s'interpréter qu'en relation avec la spécification d'un système d'oppositions. Nous en ferons, par commodité, une utilisation sensiblement plus lâche : pour nous, un phonème sera simplement un symbole ato-

mique dans un système de représentation donné de la forme phonique d'un mot. La même liberté sera prise avec les termes *phonémique* ou *phonétique*, qui seront utilisés de manière interchangeable pour qualifier ce qui est relatif à un système de transcription donné (symbole *phonétique* ou symbole *phonémique*, etc.).

Les symboles phonétiques utilisés dans les transcriptions reproduites dans ce document sont listés à la table 0.1.

Par ailleurs, dans l'ensemble du document, les notations suivantes seront respectées :

- $G = \{g_1, \dots, g_{n_G}\}$ désigne l'ensemble des symboles graphémiques (graphèmes);
- $P = \{p_1, \dots, p_{n_P}\}$ désigne l'ensemble des symboles phonémiques (phonèmes);
- $\mathcal{G}$ et $\mathcal{P}$ désignent respectivement les monoïdes libres sur G (resp. P), en d'autres termes les ensembles de toutes les chaînes de taille finie de symboles graphémiques (resp. phonétiques). λ note la chaîne vide, dans l'un comme dans l'autre alphabet. Les chaînes graphémiques apparaîtront en italique (e.g. *graphie*), et les chaînes phonémiques entre deux / (e.g. /gʁafi/, qui représente la transcription de *graphie*);
- un lexique phonétique $\mathcal{L}$ est alors un sous-ensemble de $\mathcal{G} \times \mathcal{P}$. Pour un lexique $\mathcal{L}$ donné, on notera $\mathcal{L}_G$ (resp. $\mathcal{L}_P$) le sous-ensemble de $\mathcal{G}$ (resp. $\mathcal{P}$) tel que tout élément de $\mathcal{L}_G$ (resp. $\mathcal{L}_P$) figure dans au moins un couple de $\mathcal{L}$. Nous référerons aux éléments du lexiques préférentiellement sous le terme de *lexies*, mais *entrées lexicales* ou même *mots* (!) seront parfois utilisées de manière alternative;
- une chaîne x de $\mathcal{G}$ étant donnée, on note par $\phi(x)$ une chaîne de $\mathcal{P}$ représentant sa prononciation, i.e telle que le couple $(x, \phi(x))$ soit un élément de $\mathcal{L}$. La notation fonctionnelle ne doit pas nous leurrer : si x n'est défini que comme une suite de graphèmes, alors ϕ n'est pas une application de $\mathcal{G}$ dans $\mathcal{P}$, à cause de l'ambiguïté possible de prononciation (homographes hétérophones) : ainsi $\phi(\text{couvent})$ est indéfini, puisque dans un lexique phonétique du français cohabitent les paires (*couvent*, /kuv/) et (*couvent*, /kuvā/). Si toutefois on associait à chaque chaîne un vecteur de traits morpho-syntaxiques et sémantiques, à la variabilité¹ rémanente près (c.f. les deux prononciations possibles de *ananas*), on pourrait considérer dans la pratique ϕ comme une fonction. Ceci ne sera pas le cas pour la majorité des lexiques avec lesquels nous serons amené à travailler; en conséquence il faudra lire $\phi(x)$ comme « un élément de $\mathcal{P}$, tel que $(x, \phi(x))$ est dans $\mathcal{L}$ ».

On notera également, par abus de notation, $\phi_x(s)$ (ou simplement $\phi(s)$ si il n'y a pas d'ambiguïté possible) le segment de phonèmes de $\phi(x)$ apparié

1. La question de la variabilité est une question importante et difficile sur laquelle nous reviendrons dans les pages qui suivent (page 24 et suivantes), auxquelles nous renvoyons pour une justification plus complète de notre point de vue.

Voyelles					
	[e]	fée, set		[ə]	le, above
(F)	[i]	il, vie, lyres	(A)	[i:]	bead, see
(F)	[ɑ]	pâte, bas	(A)	[ɑ:]	bard, calm
(F)	[ɛ]	laine, jouet	(A)	[ɔ:]	born, cork
(F)	[a]	la, table	(A)	[u:]	boon, fool
(F)	[ɔ]	mort, donner	(A)	[ɜ:]	burn, work
(F)	[o]	dôme, beau	(A)	[ɪ]	sit, pity
(F)	[u]	genou, roue	(A)	[æ]	sat, apple
(F)	[y]	rue, venue	(A)	[ʌ]	fun, come
(F)	[ø]	peu, deux	(A)	[ɒ]	fond, wash
(F)	[œ]	peur, meuble	(A)	[ʊ]	full, soot
(F)	[ɛ̃]	matin, plein	(A)	[eɪ]	bay, fate
(F)	[ɑ̃]	sans, vent	(A)	[aɪ]	buy, lie
(F)	[ɔ̃]	ombre, bon	(A)	[ɔɪ]	voice, boy
(F)	[œ̃]	lundi, brun	(A)	[əʊ]	no, ago
			(A)	[aʊ]	now, plough
			(A)	[ɪə]	tier, bier
			(A)	[ɛə]	tare, fair
Glides					
	[j]	filles, yet	(F)	[ɥ]	huile, lui
	[w]	oui, wet	(A)	[h]	hat, preheat
Consonnes					
	[p]	papa, pope		[l]	litre, little
	[t]	titre, strut	(A)	[r]	ran, steer
	[k]	kaki, kiss	(F)	[ʁ]	rue, venir
	[b]	bébé, baby		[m]	ram, mammoth
	[d]	dindon, daddy		[n]	ran, nut
	[g]	gogo, agog		[ŋ]	rang, bank
	[f]	fine, raffle		[v]	vin, river
	[s]	sale, rice		[z]	zéro, buzz
	[ʃ]	chance, sugar		[ʒ]	jeune, measure
(A)	[θ]	thin, math	(A)	[ð]	this, other
(A)	[tʃ]	chance	(A)	[dʒ]	just, edge
(F)	[ɲ]	agneau, vigne		[x]	jota, loch

TAB. 0.1 - Symboles phonétiques de l'alphabet phonétique international pour la transcription du français et de l'anglais.

Les symboles spécifiques au français sont marqués (F), ceux qui n'apparaissent que dans des transcriptions anglaises sont marqués (A). Sources: Le Robert & Collins, seconde édition, 1988.

à un segment s de x . Ceci suppose naturellement qu'un schème d'appariement entre les deux chaînes soit disponible;

- si x est une chaîne de $\mathcal{G}$ ou de $\mathcal{P}$, on notera $l(x)$ sa longueur, et $L(x)$ l'ensemble de ses sous-chaînes (au sens large). $L(x)$ a au plus $l(x) \times (l(x) + 1)/2 + 1$ éléments distincts. Si x et y sont deux chaînes, on notera $x \leq y$: « x est une sous-chaîne (éventuellement égale à y) de y »;
- si (x_1, x_2) est un couple de chaînes de $\mathcal{G} \times \mathcal{G}$ ou de $\mathcal{P} \times \mathcal{P}$, on note $pgcp(x_1, x_2)$ (respectivement $pgcs(x_1, x_2)$) le plus long commun préfixe (resp. suffixe) de x_1 et de x_2 .
- une *règle de transcription* graphème-phonème sera définie par la donnée de quatre termes : g le graphème (ou la chaîne de graphèmes) à transcrire, $/p/$ le phonème (ou la chaîne de phonèmes) résultat, C_g et C_r les contextes graphémiques gauche et droit de la règle. Ces règles seront représentées sous la forme $g \rightarrow p / C_g - C_r$, signifiant que g se prononce $/p/$ lorsqu'il est précédé de C_g et suivi de C_r . Par exemple, la règle décrivant la prononciation du g lorsqu'il est suivi d'un e s'écrit : $g \rightarrow \text{ʒ} / - e$. Le même formalisme sera utilisé pour décrire des règles phonologiques, à la différence près que les quatre termes sont des chaînes de phonèmes (ou des vecteurs/matrices partiellement spécifiés de traits distinctifs).

Introduction

Motivations et objectifs

Le problème de la conversion automatique d'une chaîne de caractères graphémique en sa contrepartie phonémique est un problème déjà ancien, puisque les premières réalisations en la matière remontent à plus de vingt ans. Depuis, de multiples systèmes ont été développés, leur mise en œuvre informatique a été optimisée, et des modules de transcription ont été intégrés dans de nombreux programmes commerciaux. Parallèlement, l'augmentation rapide des capacités de stockage des ordinateurs a rendu possible le développement de dictionnaires électroniques de grande dimension, incluant, pour certains d'entre eux, une ou plusieurs prononciations pour chacun des éléments lexicaux répertoriés.

Prétendre présenter aujourd'hui, dans un tel contexte, des éléments nouveaux sur cette question, ne peut que susciter la suspicion, voire l'incrédulité. C'est pourtant ce que nous allons tenter de faire dans ce document, en nous appuyant sur une expérience singulière, celle de la transcription manuelle d'un vaste lexique de noms propres. Patronymes, toponymes, et acronymes continuent en effet de poser aux systèmes de transcription automatiques des difficultés importantes, et freinent la diffusion de bon nombre de systèmes « grand-public », dont l'acceptabilité repose crucialement sur leur capacité à bien prononcer les noms propres (que l'on pense par exemple aux applications de synthèse par téléphone nécessitant que l'interlocuteur humain soit nommément identifié). Dans ce contexte, constituer un lexique enregistrant une ou des prononciations pour les noms propres les plus fréquents apporte, de manière très concrète, un début de solution à ce problème. Compte-tenu toutefois de la non-finitude de la catégorie « nom propre », et de la distribution de fréquences particulière que ces unités présentent (Sorin 1991), il est indispensable de compléter ce lexique par des mécanismes capables de transcrire les éléments qui n'y figurent pas.

Transformer une expérience de transcripteur en un savoir utilisable par des mécanismes automatiques pouvait s'envisager de deux manières différentes : le convertir explicitement sous la forme règles, ou bien envisager l'utilisation de diverses techniques d'apprentissage permettant d'effectuer automatiquement cette conversion. Ces deux chemins ont été explorés en parallèle, mais il nous a semblé, pour des raisons que nous développerons dans le corps du document, que la voie de l'apprentissage était la plus prometteuse. À ce stade de notre travail, de nombreuses alternatives étaient offertes : l'état du domaine de l'apprentissage automatique du langage naturel est tel que de nombreuses techniques permettent de

réaliser implicitement ce transfert de connaissances. Là encore, plusieurs directions de recherche ont été considérées, avant que nous soyons convaincu que les techniques à base d'analogie offrent le cadre le plus satisfaisant pour conceptualiser et formaliser un certain nombre de nos intuitions.

Pour en exprimer la principale, disons qu'elle s'inscrit dans la lignée de des travaux exposés dans (Coker, Church, & Liberman 1990), qui décrivent le module de transcription des noms propres du système de synthèse de l'anglais TTS. Comme ces auteurs le notent correctement (p. 84) :

Sometimes surprisingly simple processes provide the greatest benefits. The rhyme analogy is such one case. The pronunciation of an unknown name such as *Plotsky* is determined by analogy with *Trotsky*, which happens to be in the name dictionary. The pronunciation of *Plotsky* is computed by removing the initial /tr/ of *Trotsky* and replacing it with /pl/. It is remarkable just how many names can be pronounced in this way.

Prenant appui sur cette remarque, Coker et ses co-auteurs proposent un mécanisme de transcription par analogie, dans lequel toutefois la définition des relations de similarité est strictement exogène, c'est-à-dire qu'elle s'exprime à travers des « règles de ressemblance » qui font partie de la connaissance d'arrière-plan fournie *a priori* au système de transcription. Pour reprendre l'exemple qu'ils donnent, l'analogie entre *Plotsky* et *Trotsky* est reconnue parce que le système « sait » qu'un groupe <occlusive+liquide> peut être remplacé par un autre groupe du même type à l'initiale d'un mot sans que la prononciation de la partie terminale soit affectée.

Généralisant la démarche de Coker, Church et Liberman, l'objectif principal de ce travail est de montrer qu'une automatisation complète de l'approche analogique² est souhaitable, qu'elle n'est pas réalisée par les mécanismes d'apprentissage qui ont été mis en œuvre dans le cadre de la transcription (y compris par ceux qui se réclament explicitement de cette ligne de recherche), et qu'il est toutefois possible de proposer des formalisations alternatives de l'analogie qui permettent de fonder des algorithmes de transcription performants. Trois objectifs en un, donc : celui de justifier l'intérêt d'extraire, pour la tâche de transcription, des ressemblances du type de celle qui unit (graphiquement) *Plotsky* à *Trotsky*; celui de montrer que les techniques d'apprentissage validées sur ce problème échouent partiellement ou totalement à percevoir et à utiliser de telles similarités; celui enfin de construire un modèle alternatif de la transcription par analogie qui non-seulement reconnaisse de telles proximités, mais qui, de plus, soit capable d'en tirer profit pour produire des transcriptions correctes. À ces objectifs très généraux, nous avons adjoint une contrainte supplémentaire, qui a conduit à restreindre sensiblement notre champ d'investigation : les mécanismes de transcription à base de lexique que nous considérons doivent être adaptés à l'utilisation de corpus de très grande taille (i.e. dont les entrées se comptent en

2. Le terme *analogique* possède deux significations en français : il qualifie premièrement ce qui a trait à l'analogie, mais s'applique également aux mécanismes de transmission de données sous la forme de variation continue d'une grandeur physique. Pour éviter les confusions, précisons que c'est la première de ces deux définitions qui s'appliquera tout au long de ce document.

dizaine de milliers).

Nous avons, tout au long de cette étude, souhaité nous donner les moyens d'évaluer de manière « juste » les différentes procédures que nous avons développées, et de les comparer avec d'autres approches de ce problème. De telles évaluations et comparaisons sont notoirement délicates, dans la mesure où la précision des alphabets phonétiques utilisés dans différents systèmes est variable, et où l'exploitation de connaissances d'arrière-plan rend difficilement isolables les performances en généralisation intrinsèques d'un algorithme.

Afin de favoriser de telles comparaisons, nous avons donc été conduit à redévelopper deux systèmes proposés par d'autres auteurs, et à les tester de manière approfondie. Nous avons également cherché à définir des conditions expérimentales qui permettent de telles comparaisons; et en particulier veillé à ce que les différents algorithmes qui sont évalués travaillent exactement à partir des mêmes représentations lexicales. Dans la mesure où ces algorithmes sont inégalement capables de profiter de tel ou tel enrichissement linguistique des connaissances lexicales, nous nous en sommes tenus à des représentations « pauvres » des lexies, à savoir des couples (graphie, phonétique) alignés.

Ce parti-pris constitue un élément important de notre démarche, puisqu'il repose sur une hypothèse implicite, qui est que la simple connaissance de la représentation graphémique d'un mot permet d'en inférer la (ou les) prononciation(s). Cette hypothèse est notoirement incorrecte pour un bon nombre de langues, ce qui semble invalider *a priori* l'ensemble de notre démarche. Ce parti-pris nous paraît pourtant fondé, dans la mesure où il nous semble essentiel, au-delà du constat de l'inadéquation partiel de telles représentations pour la tâche de conversion graphème-phonème, de mesurer précisément jusqu'où il est possible d'aller avec de telles représentations, de cerner le type de problèmes que chacun des algorithmes rencontre, et d'identifier la nature des connaissances supplémentaires dont tel ou tel mécanisme d'apprentissage pourrait avantageusement tirer parti.

Une des conséquences positive de ce choix est que les techniques que nous proposons peuvent être appliquées telles que à d'autres types de problèmes mettant en jeu la transduction de séquences symboliques. Ce dernier point mérite toutefois d'être nuancé: les transductions que nous apprenons préservent les ordonnancements au sein des séquences, une condition qui n'est pas systématiquement respectée dans tous les problèmes de transduction.

Dans le même ordre d'esprit, nous aurons l'occasion de mentionner, dans le cours de la discussion, certaines « astuces » relativement générales, qui permettent, dans certaines conditions, d'améliorer les performances de généralisation. Ces astuces n'ont pas été utilisées, et, en ce sens, il est permis de dire que les performances que nous présentons sont très probablement améliorables. Ces performances représentent toutefois, à nos yeux, une mesure objective « brute » de la capacité des différentes techniques à traiter de manière satisfaisante le problème de l'apprentissage de la transcription.

Signalons finalement que toutes nos évaluations ont été effectuées sur de très gros lexiques de noms propres et de noms communs de cinq langues différentes (français, anglais, italien, allemand, néerlandais).

Organisation générale

L'organisation générale de ce document rend largement compte, au risque d'un certain désordre, du cheminement de notre réflexion, de nos essais et de nos erreurs. Il nous a en effet semblé nécessaire, pour la compréhension de l'ensemble, de ne pas effacer les traces de nos hésitations, et de restituer, le plus fidèlement possible, la manière dont l'idée des algorithmes que nous proposons s'est progressivement imposée. L'articulation des principaux objectifs que nous nous sommes assignés est néanmoins globalement respectée, puisqu'il y aura un temps pour la justification de l'approche à base de lexique (chapitre 1), un temps pour la critique de nos « prédécesseurs » (chapitres 2 et 3), et un temps pour la modélisation (chapitres 3 et 4). Cette organisation générale se détaille comme suit.

Dans un premier temps, nous introduisons les divers éléments du contexte dans lequel s'est développé notre réflexion, et justifions le choix d'une approche à base de données (et en particulier à base d'analogies) pour la transcription graphème-phonème. Nous donnerons donc, à la section 1.1, une présentation générale des différents problèmes posés par la transcription graphème-phonème, en insistant sur le problème des noms propres, dont les difficultés de traitement, nous le verrons, sont telles qu'elles rendent les approches traditionnelles (à base de règles) partiellement inopérantes. Un second élément de contexte important sera introduit avec une discussion des modèles cognitifs de la lecture, à la section 1.2. Il nous a semblé important de présenter, même sommairement, ces modèles, en raison de l'influence qu'ils ont exercé dans le développement de techniques analogiques pour la transcription. Ce tableau d'ensemble sera complété, à la section 1.3, par une présentation du concept d'analogie, et des différentes modélisations qui ont pu être proposées des mécanismes analogiques, en particulier dans le traitement du langage naturel.

Nous examinons ensuite en détail dans le chapitre 2 un certain nombre d'approches à base de lexiques de la transcription graphème-phonème. Nous avons regroupé ces approches sous le terme général d'approches *classificatoires*. Nous essaierons de montrer, en nous appuyant sur un exemple précis d'implantation (qui est présenté à l'annexe B), que ces techniques, en dépit de leur relatif succès, sont incapables de répondre correctement à un certain nombre d'exigences importantes. Ce second chapitre nous donnera ainsi des éléments supplémentaires pour justifier le choix d'une approche à base d'analogies.

Nous présenterons ensuite, au chapitre 3, trois variations algorithmiques reposant sur une même idée : celle d'une transcription par recollement de segments. Nous montrerons comment ces approches, inspirées des travaux de Dedina et Nusbaum (1991) sur la prononciation par analogie, permettent de répondre à certaines des critiques adressées aux approches classificatoires, parce qu'elles mettent en avant deux principes importants : l'abandon d'une conception « naïve » de la segmentation d'une part ; la remise en cause du primat des fréquences comme critère de décision d'autre part. En identifiant un cadre formel commun entre ces méthodes et d'autres techniques d'apprentissage, nous montrerons toutefois que les approches fondées sur des grammaires de segments manquent à prendre en compte de manière satisfaisante les effets *systématiques* résultant des relations de ressemblance/dissimilitude qui structurent en profon-

deur les lexiques.

C'est en assignant une place centrale au concept de *paradigme*, qu'au chapitre 4, nous parviendrons véritablement à intégrer la transcription par juxtaposition de segments dans une approche formelle de l'analogie. Cette nouvelle formalisation, matérialisée dans deux algorithmes, a deux conséquences particulièrement positives : d'une part un renouvellement de perspective, permettant l'abandon radical de toute de *mise en correspondance* des chaînes de graphèmes avec des chaînes de phonèmes; d'autre part la restitution algorithmique de nos intuitions de transcripateur sur l'importance des relations de similitudes dans un système appris de transcription. Nous montrerons également dans ce chapitre, en appliquant nos algorithmes à un problème de conjugaison, que les principes analogiques mis en œuvre sont également susceptibles de s'appliquer avec succès à d'autres problèmes posés par l'apprentissage automatique de systèmes de traitement du langage naturel.

Notons finalement que sont reportés en annexe un certain nombre d'éléments importants pour la compréhension générale, notamment la présentation et l'évaluation de l'algorithme classificatoire qui nous servira de référence (annexe B), de l'algorithme de Dedina et Nusbaum (annexe C), des protocoles d'évaluation (annexe D), et des lexiques utilisés pour les évaluations (annexes E à I).

Chapitre 1

Présentation du contexte

Nous introduisons dans cette section trois éléments importants du contexte dans lequel se sont effectués nos travaux. Dans un premier temps, nous présentons de manière générale le problème de la transcription graphème-phonème (section 1.1), en développant un certain nombre de réflexions personnelles qui motivent l'orientation qui a été prise, à savoir le traitement de ce problème par des techniques d'apprentissage.

Nous présentons ensuite une introduction sommaire aux modèles cognitifs de la lecture (à la section 1.2). Ce second élément de contexte est probablement moins essentiel à la bonne compréhension du document que le précédent, dans la mesure où, dans l'ensemble, la plausibilité cognitive des modèles que nous avons évalués ou développés n'a pas été, à proprement parler, au centre de nos préoccupations. Toutefois, il nous a semblé nécessaire d'accorder une place, même réduite, à une telle présentation, dans la mesure où ces modèles cognitifs ont inspiré et motivé les premières réalisations informatiques de systèmes de prononciation à base d'analogies, et qu'à ce titre, ils sont récurrentement invoqués pour justifier la pertinence de ces approches. Ils constituent donc un élément d'arrière-plan qui permet de mieux comprendre un certain nombre d'éléments du débat, qui risqueraient, sans cela, de rester obscurs.

Nous concluons ce chapitre en discutant du concept d'analogie (voir la section 1.3), en insistant d'une part sur la formalisation qui en est donnée dans le domaine de l'intelligence artificielle, et d'autre part sur quelques tentatives de mise en œuvre dans le domaine du traitement du langage naturel. En particulier, cette section terminale nous donnera l'occasion de mettre en évidence deux conceptions relativement différentes de l'analogie.

1.1 La transcription graphème-phonème

Nous posons dans cette section un premier élément essentiel du contexte dans lequel nous avons effectué nos travaux, en présentant le problème de la transcription graphème-phonème. Pour l'essentiel, cette présentation consistera en un rapide survol des difficultés que pose l'automatisation de cette tâche (voir

la section 1.1.2), ainsi que de quelques-unes des (nombreuses) réalisations dans ce domaine (à la section 1.1.3).

Notre seule contribution réellement originale à cette section concerne le traitement des termes extra-lexicaux (principalement les noms propres), problème que nous avons été amené à envisager et traiter de manière relativement complète (voir sur ce point la section 1.1.2). Cet aspect spécifique sera l'occasion de mettre en évidence certaines limitations des approches à base de règles d'experts, et de montrer l'utilité d'une approche à base de données.

1.1.1 Présentation générale

La tâche à résoudre

Parmi les différents systèmes de notation de la parole, institués par les sociétés humaines au cours de leur développement, on distingue (Llorach 1968) trois types principaux :

- les systèmes *idéographiques* ou *logographiques*, dans lesquels chaque symbole graphique représente un contenu signifié différent. Le système d'écriture le plus conforme à ce principe, en dépit de déviations importantes qui le rapprochent sous divers aspects des systèmes sémiographiques, est le système d'écriture du chinois;
- les systèmes *sémiographiques*, dans lesquels chaque symbole graphique représente un signe (i.e. un signifiant *et* un signifié). Ce système est pour le langage, extrêmement coûteux en symboles;
- les systèmes *phonographiques*, dans lesquels idéalement chaque symbole graphique (chaque graphème) représente un son différent du langage. Dans ces systèmes, c'est en fait le signifiant qui est représenté graphiquement, ce qui, compte-tenu de la petite taille des inventaires phonémiques, rend ces systèmes plus économiques. Selon les langues, divers niveaux de notation ont été adoptés (syllabique, consonantique, et alphabétique). Des exemples de tels systèmes d'écriture sont le japonais (notation syllabique), les langues sémitiques (notation consonantique), et les langues romanes (notation alphabétique).

Ceci étant posé, le problème de la conversion graphème-phonème consiste très précisément à reconstituer, à partir d'une forme graphique (un mot, un syntagme, une phrase) exprimée dans un système phonographique (nous ne considérerons par la suite que des systèmes alphabétiques), la forme orale qui lui correspond. Cette forme orale est exprimée sous l'aspect d'une séquence de phonèmes. Selon les langues, cette tâche présente des degrés de difficulté variables, le français se situant probablement parmi celles qui posent le plus de problèmes, ainsi que nous allons essayer d'en rendre compte dans les paragraphes qui suivent. Pour l'essentiel, cette discussion des difficultés de transcription portera sur la reconstitution de la chaîne segmentale. N'oublions pas toutefois que, pour bon nombre d'applications, une transcription complète doit également comporter des informations prosodiques : localisation des accents, la

longueur des phonèmes, ou nature des tons associés aux unités phonétiques. De telles informations sont requises, pour bon nombre de langues, y compris lorsque l'on ne se préoccupe que de la transcription de mots isolés.

Les applications des systèmes de transcription

Un certain nombre de motivations différentes peuvent justifier le développement d'un système automatique de transcription graphème-phonème. Parmi les applications les plus directes d'un tel système, mentionnons principalement :

– La description linguistique

L'étude de la correspondance entre écrit et oral, de son histoire, des nombreux types d'ambiguïté qu'elle recèle (Catach 1984), et son expression complète dans un système de règles, constitue en soi un sujet d'études, qui permet d'une part de mieux comprendre les mécanismes d'évolution de la langue, et d'autre part d'identifier les unités orthographiques qui sont réellement pertinentes pour la lecture, avec un champ d'application immédiat : la didactique du français.

– Le traitement automatique de la parole

- Synthèse de la parole. Dans le cadre de la synthèse de la parole à partir du texte, une étape indispensable du traitement consiste à fabriquer la chaîne segmentale qui sera synthétisée. C'est précisément autour de cette problématique que nous avons été conduit à aborder la question de la transcription;
- Reconnaissance de la parole. Les systèmes de reconnaissance de la parole, pour être capables de traiter de larges vocabulaires, sont, dans la pratique, contraints de procéder à une reconnaissance globalement phonémique¹ (Bartkova, Larreur, & Métayer 1994). Ces systèmes requièrent donc que des transcriptions phonétiques de chacun des mots du vocabulaire à reconnaître soient disponibles, transcriptions dont il est tout spécialement important qu'elles représentent explicitement les diverses réalisations possibles. Le développement d'applications grand-public, dans lesquelles l'utilisateur pourra augmenter le vocabulaire reconnu, en ajoutant des formes *graphémiques*, impose l'intégration dans les systèmes de reconnaissance d'outils performants de transcription.

– Le traitement automatique de la langue écrite

Dans le domaine du traitement de l'écrit, le besoin d'un système de transcription se fait principalement sentir pour des applications de correction orthographique. En effet, pouvoir retrouver la forme correcte d'un mot ou d'un nom incorrectement orthographié nécessite dans bien des cas le

1. Les unités effectivement reconnues peuvent, dans la pratique, être différentes du phonème (pseudo-diphone, allophone), mais la sortie globale du système de reconnaissance est phonémique.

passage par une représentation phonémique (ou approximativement phonémique) (Pecatte 1992). L'idée générale du traitement est la suivante : la forme inconnue (i.e. potentiellement incorrecte) est transcrite, et la transcription est comparée à celle des mots du dictionnaire. Un calcul de distance, intégrant les différences graphiques et phonétiques, permet finalement d'inférer le mot qui a été mal orthographié. Un mécanisme équivalent est également à l'œuvre dans l'interface d'interrogation du minitel, ce qui la rend relativement robuste à des erreurs dans l'orthographe des noms.

Dans le domaine de l'écrit toujours, on notera un second domaine d'application, bien plus marginal, celui de l'étude statistique des structures métriques en poésie (Beaudoin & Yvon 1994; 1996).

Signalons finalement que, suivant les applications, le problème de la transcription ne se posera pas tout à fait dans les mêmes termes. Dans un cas (celui de la synthèse vocale, par exemple), on souhaitera produire une transcription unique, dans un autre (reconnaissance), on aura besoin de générer plusieurs variantes, dans un autre encore (correction orthographique) on utilisera un inventaire phonémique moins complet, etc.

1.1.2 Les difficultés de la transcription graphème-phonème

Nous présentons, dans les paragraphes qui suivent, un inventaire des difficultés que pose la transcription automatique d'une suite de graphèmes en une suite de phonèmes. Nous ne discuterons pas ici des problèmes (difficiles) qui peuvent se poser en amont de la transcription proprement dite, à savoir les problèmes de segmentation du texte, d'identification des unités lexicales, et d'étiquetage morpho-syntaxique.

Nous aborderons tout d'abord deux points très généraux, d'une part la question de la variabilité, et d'autre part la question de l'écriture de l'oral. Nous détaillerons ensuite les problèmes spécifiques à divers traitements particuliers : la transcription des éléments du vocabulaire commun (éléments lexicaux ou para-lexicaux), la transcription des noms propres (éléments extra-lexicaux), la transcription de séquences excédant le mot.

La question de la variabilité

Le premier problème très général qui se pose est celui de l'instabilité diachronique et de la variabilité synchronique des formes orales. Ce dernier point a été étudié en profondeur pour le français par Lacheret-Dujour (1990), qui identifie, au niveau phonologique, deux types principaux de variabilité :

- une variabilité *systématique*, conditionnée par des facteurs de nature linguistique, au premier rang desquels les contextes phonémique et grammatical. Ce type de variabilité est illustré par les trois réalisations différentes

du mot *six* dans les trois occurrences suivantes :

- Luc a six (/si/) billes;
 - Luc a six (/siz/) ans;
 - Luc en a six (/sis/);
- une variabilité « *libre* », c'est-à-dire conditionnée par des facteurs extra-linguistiques. Entrent dans ce cadre :
- les variantes régionales : la coexistence, au sein d'une communauté de locuteurs d'une même langue, de plusieurs systèmes phonologiques légèrement différents. L'exemple des locuteurs du français est, à cet égard, caractéristique : entre un méridional pour lequel les oppositions entre certaines voyelles (/e/ et /ɛ/, par exemple) sont neutralisées, et un locuteur parisien pour lequel elles restent distinctives, les situations de malentendus ne manquent pas;
 - les variantes idiolectales, dues à l'utilisation des codes linguistiques, et en particulier du code oral, comme un système de marquage social (par âge, classe sociale, etc.), favorisant l'identification des pairs au sein d'un groupe, et permettant au contraire la stigmatisation de ceux dont l'utilisation du code est défailante par rapport aux normes socialement instituées. Les emprunts à des langues étrangères, dont le flux ne cesse de déborder les institutions de normalisation des prononciations, constitue à cet égard un *locus* de variabilité particulièrement riche, parce qu'il donne au groupe dominant l'occasion de réaffirmer, par l'expression de sa familiarité avec une ou des langues étrangères, sa suprématie culturelle;
 - des variantes prosodiques, liées au contexte d'élocution au sens large, et entraînant par exemple des variations dans la réalisation des liaisons, ou, en relation à une accélération du débit, des variations dans la réalisation de diérèse, dans le nombre de /ə/ élidés, ou d'assimilations réalisées.

L'existence de ces nombreux facteurs de variabilité est une première entorse au principe de correspondance bi-univoque entre séquences orthographiques et phonétiques : cette correspondance n'est pas une bijection, ce n'est même pas, en toute rigueur, une application. La transcription graphème-phonème n'a donc de sens qu'en référence à une norme de prononciation donnée, qui spécifie à la fois les alphabets de symboles valides en entrée comme en sortie, et le niveau stylistique de prononciation que ces symboles encodent.

Écrire la parole

Toujours en restant sur un plan très général, on notera également que dès ses prémisses, le projet de reproduire fidèlement la forme orale par une suite de symboles graphiques était compromis. Ainsi les grecs, grands précurseurs en matière de notation alphabétique, ne marquaient pas graphiquement la longueur de toutes les voyelles.

En effet, les formes graphiques ne notent en général pas les phénomènes complètement prédictibles pour un locuteur compétent d'une langue. Un exemple caractéristique de cet état de fait est en français la notation graphique d'un mot comme *absolument*. L'assimilation régressive de la surdité du /s/ entraîne quasi-mécaniquement la réalisation du *b* en /p/. Le caractère automatique de cette assimilation rend sa notation graphique superflue.

Les éléments lexicaux ou para-lexicaux

Si l'on fait abstraction des remarques précédentes, les difficultés que peut rencontrer un système de transcription automatique des entrées lexicales sont dues à une double déviation du système orthographique du français contemporain par rapport à un système de notation idéal (dont se rapproche une langue telle que l'espagnol).

La correspondance "une lettre"-"un son" n'est tout d'abord pas complètement respectée, puisqu'à un son peuvent correspondre plusieurs graphèmes (ainsi *ch* note le son /ʃ/), et qu'à une lettre peuvent correspondre plusieurs sons (ainsi *x* note dans *axe* les sons /ks/). Ceci génère des problèmes de segmentation, plusieurs découpages d'une chaîne pouvant entrer en concurrence. Considérons par exemple la forme *gageure* : la séquence *geu* peut s'interpréter de deux manières différentes, soit en *ge+u*, entraînant la prononciation /ʒu/, soit en *g+eu*, entraînant la prononciation ʒœ. Cette règle de correspondance bi-univoque est d'ailleurs d'autant moins bien respectée qu'en français, le même graphème (ou groupe de lettres) peut représenter, suivant le contexte où il figure, plusieurs phonèmes, et un phonème (ou groupe de phonèmes) être représenté par plusieurs lettres. Pour donner un exemple, le graphème *s* se prononce en général /z/ lorsqu'il apparaît entre deux voyelles, et /s/ sinon.

Un second problème important est posé par l'existence d'homographes hétérophones, c'est-à-dire de formes qui partagent une même représentation graphique, mais dont la prononciation (et généralement le sens) sont différents. Les exemples de tels ambiguïtés sont bien connus (*violent:Adj* et *violent:Vb* en est un), comme le sont, en principe, les moyens pour les résoudre (identifier au préalable, les catégories morpho-syntaxiques des mots à transcrire, voire, dans certains cas, procéder à une désambiguïsation sémantique).

Plusieurs raisons peuvent être invoquées pour expliquer ces déviations par rapport à une notation non-équivoque. Nous en mentionnerons brièvement trois :

- la langue évolue, et l'évolution du parler est bien plus rapide que celle de l'écrit, dont les changements sont soumis à l'inertie des instances de contrôle et de normalisation. En conséquence, les deux systèmes de représentation ne cessent d'être en décalage, la forme écrite conservant des traces de prononciations obsolètes. Un exemple de ce phénomène est l'évolution de la prononciation du digraphe *oi*, dont la prononciation est progressivement passée de /ɔi/ à /ue/, puis à /we/ et finalement /wa/, sans que l'orthographe enregistre systématiquement ces changements;
- la langue évolue aussi en empruntant à d'autres communautés linguistiques, au fur et à mesure des besoins ou des circonstances historiques, des

formes nouvelles. Ceci engendre inmanquablement des incohérences supplémentaires dans le système de correspondance, qui doit rendre compte d'appariements graphème-phonème empruntés à d'autres systèmes phonographiques;

- plus profondément, l'institution d'un second type de signifiant (le signifiant graphique), même conçue dans le cadre d'un assujettissement complet à un autre signifiant (le signifiant oral), ne peut, diachroniquement, que conduire à l'apparition d'oppositions et de distinctions qui ne sont pertinentes qu'au sein du nouveau système sémiotique ainsi constitué. Les manifestations de cette autonomisation partielle du système graphémique sont nombreuses, et resurgissent lors de chaque débat sur la réforme de l'orthographe, à travers l'argument de « sauvegarde des étymologies ». Pour ne donner que deux exemples de ces oppositions purement graphiques, mentionnons, sur le plan morpho-syntaxique les oppositions entre les différentes terminaisons verbales du français (*ais, ait, aient*), qui ne sont pas marquées à l'oral (sinon à travers les liaisons). Mentionnons également la multiplicité des oppositions graphiques correspondant aux finales en /o/ (*o, ot, eau, au, aud...*), qui ne se justifient qu'au regard de la nécessité de conserver graphiquement des distinctions qui signalent soit l'origine de la forme, soit son inscription dans un paradigme dérivationnel donné (le *e* de *agneau* signale ainsi l'alternance /o-ə/ dans la dérivation *agneau/agneler*); soit encore son appartenance à une classe d'objets (*eau* est une finale caractéristique des petits d'animaux (*renardeau, baleineau, etc.*)) (Gruaz 1995).

Les contraintes internes auxquelles est ainsi soumis le système orthographique contribuent à rajouter de la confusion dans le système de correspondance entre graphie et phonétique, à un point tel qu'il est légitime, à notre sens, de questionner la validité du point de vue qui consiste à considérer le système orthographique comme assujetti aux formes orales, et de lui reconnaître une dynamique propre.

Les éléments extra-lexicaux

Un certain nombre de difficultés supplémentaires sont causées par la présence dans les textes à transcrire d'éléments qui ne font pas partie, par principe, des lexiques de langue. Pour une large majorité, ces éléments correspondent à des noms propres ou à des dérivés de noms propres. La proportion de ces noms propres est évidemment très variable suivant les textes, mais elle peut atteindre, pour des applications de lecture spécifiques (applications de synthèse par téléphone nécessitant une identification nominative du locuteur) ou des types de textes particuliers (articles de journaux), des niveaux relativement importants, trop importants pour que le traitement de ces éléments particuliers soient négligés.

Parmi les noms propres, nous distinguerons entre deux types principaux, qui posent des problèmes assez différents, à savoir d'un côté les noms de personnes et de lieux, de l'autre les groupes nominaux complexes soumis à des mécanismes de siglaison. Ces composés correspondent à des noms de groupes d'individus ou

d'intérêts publics et privés, ou encore à des termes du vocabulaire administratif, technique ou scientifique.

On notera qu'en la matière, la situation des différentes langues est bien plus égalitaire que pour le vocabulaire commun, puisque ces problèmes se posent pratiquement partout à l'identique (voir par exemple (Carlson, Granström, & Lindström 1989; Hochberg *et al.* 1990; Spiegel & Machi 1990; Vitale 1991; Golding 1991; van Coile, Leys, & Mortier 1992; Gaved 1993; Belhoula 1993) pour le traitement des noms propres dans diverses autres langues, et (McCully & Holmes 1988; Diamond 1994; Viana, Trancoso, & Silva 1994) pour le traitement des sigles et des noms de société).

Les noms de personnes et de lieux Nous avons regroupé dans un même chapeau patronymes et toponymes, dont la transcription automatique pose des difficultés sensiblement similaires. On peut identifier grossièrement trois sources majeures de difficultés (Métayer & Vincent 91; Yvon 1993) :

- **L'absence d'autorité instituée de normalisation des graphies.** À la différence des noms communs, pour lesquels diverses institutions spécifiques sont officiellement (l'Académie) ou officieusement (les éditeurs des principaux dictionnaires) mobilisées pour maintenir l'ordre et la cohérence de l'organisation du système orthographique, il n'existe aucune autorité équivalente pour les noms propres, ce qui justifie le lieu commun « les noms propres n'ont pas d'orthographe ». Historiquement, les noms propres (toponymes et patronymes) français ont évolué relativement librement depuis le bas-français ou le latin, et conservent, en maints aspects, un bon nombre de séquelles orthographiques qui ont été, depuis le XVII^e siècle, progressivement éliminées des noms communs. Citons, par exemple, le maintien du *s* archaïque, remplacé presque partout dans l'orthographe par un accent (circonflexe ou aigu), mais que l'on trouve encore dans *Fresne* ou *Le Chesnay*; le maintien dans l'orthographe (voire dans la prononciation !) de *l* dans les groupes *aulC*², dont la vocalisation en *u* est pourtant ancienne (ainsi *Aulnay*, *Saulnier*, etc.).

L'absence de norme a également favorisé le foisonnement de nombreuses variantes orthographiques pour un même nom, générant en surface une diversité disproportionnée par rapport au nombre des étymologies. Pour donner un seul exemple de cette prolifération, considérons avec (Dauzat & Rostaing 1963, p. 20) le toponyme *Annay*, formé du latin *alnus* (*aulne*) et du suffixe *-ētum* (que l'on retrouve dans *arboretum*). Pour cette même étymologie, les variantes suivantes : *Annois*, *Aulnay*, *Aunay*, *Aulnois* et *Aulnoy* représentent toutes des toponymes attestés. Le même phénomène s'est produit, à une échelle plus grande encore, pour les patronymes (voir sur ce point la section 4.2.2).

Le résultat concret, du point de vue de la transcription graphème-phonème, de ce type de fantaisies orthographiques est, d'une part, une forte augmentation de l'ambiguïté de certaines séquences de graphèmes (ainsi *as*, *es*, *is* deviennent ambigus dans des contextes où, pour les

2. Conformément à l'usage établi, nous notons *C* une consonne (graphique ou phonétique) quelconque et *V* une voyelle quelconque.

mots du lexiques, ils ne le sont pas); et d'autre part par l'apparition de phonogrammes qui ne sont pas (quasiment) rencontrés dans le vocabulaire commun: par exemple les séquences finales *eix*, *etz* et *eim*, ou encore la séquence initiale *ll* (pour les noms catalans), etc.

- **L'hétérogénéité des origines.** Les lexiques de noms propres français se caractérisent, en second lieu, par une très grande hétérogénéité vis-à-vis des langues (y compris patois, dialectes) auxquelles les lexies se rattachent. Ce fait est la conséquence historique d'un double mouvement: d'une part le rattachement progressif à la couronne, puis à l'état français d'un certain nombre de territoires initialement situés dans des zones linguistiques variées, où l'usage du français ne s'imposa que tardivement (la Bretagne, le Midi-Pyrénées, la Savoie, la Corse, l'Alsace, etc.); d'autre part le recours massif, dans les régions industrielles, à la main d'œuvre d'origine étrangère, engendrant une immigration importante en provenance de diverses régions d'Europe (Pologne, Italie, Portugal), ou d'Afrique du Nord. À ces deux tendances lourdes, diverses causes mineures (le rôle passé de la France comme terre d'accueil de populations opprimées, la décolonisation, etc.) se sont sur-ajoutées pour faire du stock de toponymes et de patronymes français une véritable mosaïque linguistique.

Les problèmes concrets que cette situation engendre sont extrêmement plus complexes à résoudre que les problèmes de normalisation, en particulier parce que l'ambiguïté créée par la coexistence de formes graphiques extraites de systèmes orthographiques et phonétiques différents, est particulièrement difficile à démêler si l'origine du patronyme ou toponyme n'est pas connue avec précision. Donnons un exemple de cas particulièrement symptomatique ((Fouché 1959) est une source quasi inépuisable d'exemples): les noms de villes se terminant par *ens* sont prononcés /ɛns/ dans le domaine de langue d'Oc (*Puymorens*, *Saint-Gaudens*...), mais /ɑ̃/ dans la partie Nord (*Confolens*, *Forens*...), sans oublier ceux pour lesquels les deux prononciations existent, et quelques noms en /ãs/ (*Lens*, *Sens*). De telles difficultés se retrouvent au niveau des patronymes, avec des graphèmes ou des séquences de graphèmes extrêmement ambiguës, au premier rang desquelles le *e* (/e/ dans *Adelino*, mais /ə/ dans *Adeline*); le suffixe *er* (/e/ dans *Boucher*, et /œʁ/ ou /ɛʁ/ dans *Kircher*), et bien d'autres du même type.

À origine connue, toutefois, la difficulté de transcription de ces emprunts particuliers est très variable. En fait, elle est inversement proportionnelle à la proximité entre les langues sources et le français. Pour les langues qui utilisent un système d'écriture différent de l'alphabet romain, l'utilisation d'un système de translittération fondé sur la phonétique, pour autant qu'il soit respectueux des règles de prononciation du français (i.e. système de transcription de l'École Française d'Extrême-Orient (E.F.E.O) vs pinyin pour le chinois), conduit à des formes graphiques facilement prononçables: c'est le cas des noms arabes, chinois ou vietnamiens. Les noms propres d'origine européenne (Maes 1993) posent eux des problèmes sensiblement plus compliqués, pour au moins deux raisons. La première est que certaines langues, par exemple les langues nordiques et certaines langues slaves, utilisent un système alphabétique légèrement différent du nôtre, qui contient en particulier un certain nombre de diacritiques qui

n'existent pas en français. Ces diacritiques sont en général improprement assimilés au caractère de l'alphabet français le plus proche, quand bien même leur prononciation habituelle est sans rapport avec celle de la lettre française qui les remplace : ainsi les *ł* (/w/), *ó* (/u/), ou *ć* (/tsʲ/) polonais, le *č* (/ʃ/) tchèque, ou encore les *ą* (/ɔ/) et *ø* (/ø/) danois, etc.

Une seconde raison est que le degré d'exposition aux langues européennes d'un locuteur du français est, en général, plus variable que pour d'autres langues plus exotiques, conduisant à une variabilité plus forte des prononciations. Un système de transcription des noms propres se doit donc d'aborder de front le problème de la multiplicité des transcriptions acceptables, y compris à niveau stylistique de transcription fixé.

– **Problèmes de segmentation.** Les toponymes et patronymes posent finalement un certain nombre de problèmes de segmentation, liés à leur structure morphologique. Un exemple caractéristique est celui des noms propres français qui commencent par la séquence *Les*. Trois cas de figure sont effet possibles, qui entraînent des prononciations différentes :

- le nom est formé par agglutination du déterminant *le*, sous sa forme élidée, à une racine commençant par *es*. Cette configuration est toujours observée dans un contexte où une consonne suit le *s*. La prononciation est donc /ləs/, comme dans *Lestrade*, sauf naturellement si le *s* est archaïque, conduisant à la prononciation /le/, comme dans *Leschevin*;
- le nom est formée par agglutination du déterminant *le* à une racine commençant par un *s*. La prononciation correcte est conséquemment /ləs/, comme dans *Lesage* ou *Lesourd*;
- le nom est formé par agglutination du déterminant *les* à une racine. Selon le contexte graphémique qui suit le *s*, la prononciation appropriée est alors /le/ (*Lesvignes*) ou /lez/ (*Lesabeilles*).

De tels cas d'ambiguïté sur la décomposition morphologique ne sont pas rares, et se produisent avec un certain nombre d'autres préfixes ou suffixes. Mentionnons par exemple les deux formes de *mons* (montagne), que l'on retrouve d'une part dans *Montrouge* (de *Mont+rouge*), où le *t* est muet, et dans *Montrésor* (de *Mons+trésor*), où le *t* est prononcé; ou encore les noms commençant par *des*, etc. On notera finalement que l'hétérogénéité linguistique des formes, évoquée ci-dessus, est un facteur supplémentaire de complication, puisque certaines séquences sont interprétables comme suffixes pour certaines langues, et pas pour d'autres : ainsi *ova* s'interprète comme un suffixe tchèque dans bien des cas, mais pas dans *Casanova*; *son* est un suffixe nordique, sauf dans des noms comme *Lusson*, etc.

En résumé, les difficultés de la transcription graphème-phonème des patronymes et toponymes sont sensiblement les mêmes que ceux que l'on rencontre pour les éléments du vocabulaire commun, à une échelle différente toutefois, puisque les emprunts comme les idiosyncrasies liées à des orthographes aberrantes se rencontrent dans des proportions incommensurablement supérieures, multipliant en particulier les possibilités de variation libre.

Les sigles Les noms de groupes d'individus ou de groupes d'intérêts se présentent généralement sous la forme de groupes nominaux relativement complexes et longs (*Association des pêcheurs à la ligne de l'Aveyron*, *Confédération Française Démocratique du Travail*), auxquels il est particulièrement économique de substituer une forme abrégée. Cette forme abrégée est généralement construite en concaténant les initiales de tout ou partie (les prépositions et déterminants qui figurent dans le groupe nominal peuvent être omises) des constituants du nom composé (APLA³, CFDT⁴).

Il existe en fait une autre manière de créer des sigles, qui consiste à sélectionner d'abord la forme abrégée, choisie homographe à une lexie (nom propre ou commun) connue, puis à construire la forme complète (Mortureux 1994). Dans ce cas, la contrainte de sélection des initiales est affaiblie, et d'autres lettres de la forme développée peuvent être mises à contribution, comme dans REMEDE⁵. Cette stratégie de construction vise à créer des associations sémantiques entre l'entité désignée par le sigle, et la forme homonyme. Comme le note Mortureux, (1994, p15), à propos de l'homonymie avec un nom propre :

(...), la motivation qui repose sur une homonymie avec un nom propre suppose une antonomase de ce nom, lui permettant d'acquérir un signifié : l'aptitude au guidage pour ARIANE⁶, la rapidité (du vent) pour EOLE⁷ (...)

Ces acronymes ne posent, à condition qu'on soit capable de les repérer (ainsi, la concordance graphique entre CAPES⁸ et la forme pluriel de cape est purement fortuite), pas de difficulté de transcription spécifique : on saura les prononcer si on sait prononcer la lexie correspondante. Nous nous en tiendrons donc dans la suite aux sigles réellement construits par abréviation.

Ces sigles posent deux types de problèmes au systèmes de transcription graphème-phonème :

- **Le choix d'une stratégie d'oralisation.** L'observation n'est pas nouvelle, il existe différentes manières de prononcer un sigle, les deux principales étant la lecture (e.g. DEST⁹ lu /dest/) ou l'épellation (e.g. ENST¹⁰ lu /œneste/). Des stratégies alternatives sont, plus rarement, observées : oralisation « mixte », mêlant lecture et épellation (e.g. FN-CASL¹¹ lu /ɛfenkasel/), ou encore utilisation des procédures d'épellation non-standard (e.g. IEEE¹² lu /itrwazø/).

Les travaux de Marc Plénat (1991; 1992; 1994) ont mis en évidence un certain nombre de contraintes linguistiques, calculables à partir de la forme

3. Association des Pêcheurs à la Ligne de l'Aveyron

4. Confédération Française Démocratique du Travail

5. RELais MÉDICAL aux DÉlaissés

6. Auxiliaire de Recherche d'Itinéraire par Affichage NumérisÉ

7. Est-Ouest Liaison Express

8. Certificat d'Aptitude Pédagogique à l'Enseignement Secondaire

9. Direction de l'Enseignement Supérieur des Télécommunications

10. École Nationale Supérieure des Télécommunications

11. Fédération Nationale des Commerces d'Articles de Sport et de Loisirs

12. Institute of Electrical and Electronics Engineers

du sigle, qui pèsent de manière concurrente et parfois contradictoire sur le choix d'un mode d'oralisation. Ces contraintes sont les suivantes :

1. la *contrainte de grammaticalité*, qui impose qu'un sigle soit analysable en syllabes phonotactiquement légales. Cette contrainte force l'épellation de tout ou partie du sigle;
2. la *contrainte de poids*, qui identifie un poids¹³ critique minimal pour un mot du français; cette contrainte impose qu'un sigle se compose d'au moins deux voyelles non-consécutives (patron VCV, VCCV, etc), ou d'au moins une syllabe fermée (patron CVC), pour être lu. Cette contrainte force l'épellation de la quasi-totalité des sigles de type CV, tels que GO¹⁴ ou SO¹⁵, dont la forme lue n'est pas assez « lourde »;
3. la *contrainte d'attaque*, qui impose qu'un sigle lu contienne au moins une attaque consonantique (i.e. une séquence CV). Cette contrainte permet de justifier l'épellation des sigles de type VC ou VCC, tels que AP¹⁶ ou ACP¹⁷;
4. la *contrainte de longueur*, qui tend à interdire, chaque fois que cela est possible, l'épellation des sigles de quatre lettres et plus;
5. la *contrainte de prohibition du hiatus*, qui agit dans le sens d'une réduction du nombre de voyelles consécutives *dans les formes lues*¹⁸. Cette contrainte permet, dans une certaine mesure, d'expliquer le faible taux de lecture des sigles en CVV, et plus généralement, les plus faibles taux de lectures des sigles comportant des suites de voyelles, par rapport à ceux qui présentent une alternance régulière de consonnes et de voyelles.

Il existe naturellement d'autres facteurs qui favorisent l'un ou l'autre mode d'oralisation, qu'ils soient linguistiques (degré de lexicalisation) ou extralinguistiques (la sélection de la forme lue manifeste une plus grande familiarité avec l'entité désignée, ou exprime un jugement dépréciatif (ainsi RATP²¹, parfois prononcée /ʁtap/)). Toutefois ces remarques sont suffisamment précises pour constituer les fondements d'une mise en œuvre algorithmique (Boula de Mareuil 1994; Yvo).

- **La transcription proprement dite.** Une fois le mode d'oralisation déterminé, se pose le problème de la construction d'une prononciation. Nous

13. Plénat énonce cette contrainte en mesurant le poids d'une forme en nombre de *mores*. Nous donnons ici une version simplifiée de cette contrainte, pour ne pas alourdir exagérément notre exposé.

14. Gentil Organisateur

15. Service d'Ordre

16. Assistance Publique

17. Agence Centrale de Presse

18. Il semble que certaines contraintes ne pèsent pas de même manière sur les formes épellées et les formes lues. En effet, la forme épellée fonctionne phonologiquement comme un mot-composé, avec une frontière relativement nette entre les parties correspondant à chacune des lettres. En témoignent la synérèse dans les sigles tels que IRTS¹⁹ (/iʁʁʁʁʁʁ/ et non */iʁʁʁʁʁʁ/), l'impossibilité d'appliquer les règles d'harmonisation vocalique dans FFI²⁰ (/ɛfɛfi/ et non */ɛfɛfi/), etc (Plénat 1994, p.154)

21. Régie Autonome des Transports Parisiens

avons montré, dans (Yvon 1994a), que les règles de prononciation des sigles diffèrent en un certain nombre de points des règles de prononciation standard, en particulier en ce qui concerne la lecture des digraphes vocaliques et la nasalisation des voyelles. Ainsi, alors que *AU* et *OU* conservent leur valeur (respectivement /o/ et /u/), *AI* (e.g. MAIF²²), *EI* (e.g. SEITA²³), ainsi que probablement *OI*, sont systématiquement lus en prononçant distinctement les deux voyelles. Ainsi également, alors que *AN* et *ON* se lisent dans les sigles /ā/ et /ō/ respectivement, les voyelles *E* (e.g. ENSI²⁴) et *I* (e.g. INRA²⁵) suivies de *N* ne sont généralement pas nasalisées.

Nous avons proposé et validé une explication à ces déviations, en suggérant qu'à travers les règles de transcription une contrainte additionnelle, baptisée *contrainte d'adéquation*, est à l'œuvre, qui impose que la forme oralisée permette de retrouver, de manière aussi peu ambiguë que possible, la forme graphique correspondante.

Si la voie semble relativement dégagée pour ce qui concerne la lecture des sigles, il nous semble toutefois qu'un effort de formalisation est encore nécessaire pour parvenir à une description complètement satisfaisante du problème. Cette formalisation passe, à notre sens, par l'abandon d'un système de traitement linéaire unidirectionnel, dans lequel les sigles sont d'abord catégorisés (entre ceux qui sont lus et ceux qui sont épelés) par un système de règles implantant les contraintes énoncées ci-dessus, puis transcrits par un second système de règles.

Nous pensons, en effet, que la mise en œuvre d'une approche explicitement centrée sur la satisfaction de contraintes offrirait un cadre descriptif et algorithmique mieux adapté à la prononciation des sigles. Cette nouvelle approche, à l'imitation du traitement préconisé dans la théorie de l'optimalité (Prince & Smolensky 1992), procède en deux étapes distinctes :

- génération de *toutes* les transcriptions possibles du sigle, incluant plusieurs lectures différentes dans les cas d'ambiguïté, plusieurs épellations dans les cas où des lettres sont répétées, et ainsi que les formes correspondant à des stratégies combinant lecture et épellation;
- classement des formes au regard de leur degré de satisfaction des diverses contraintes identifiées;

Cette reformulation présente l'avantage de résoudre élégamment le paradoxe posé par la formulation des contraintes de Plénat : ces contraintes, qui censément orientent le choix d'un mode d'oralisation, portent sur la forme phonétique du sigle, c'est-à-dire nécessitent pour leur calcul que la forme phonologique soit déjà connue. Cette reformulation présente, en outre, l'avantage d'offrir un cadre unifié pour exprimer les contraintes qui pèsent sur le choix d'un mode d'oralisation, et celles qui pèsent sur le choix de règles de transcription, permettant une meilleure prévision des formes apparemment « déviantes ».

22. Mutuelle Assurance des Instituteurs de France

23. Société Nationale d'exploitation industrielle des tabacs et allumettes

24. Écoles Nationales Supérieures d'Ingénieurs

25. Institut National de Recherche Agronomique

Considérons à titre d'exemple le sigle ASCII²⁶, et quelques unes des prononciations²⁷ possibles :

1. /assi/
2. /asi/
3. /aski/
4. /æssedøzi/
5. /æsseii/

Dans le cadre pour lequel nous plaidons, la forme (3) est effectivement prédictible, car (1) comme (2) représentent trop imparfaitement la forme graphique (violation de la contrainte d'adéquation), et (4) et (5) sont trop longues (violation de la contrainte de longueur, sans parler des problèmes d'hiatus).

Un travail important reste à effectuer pour compléter la modélisation de ce traitement, en particulier pour trouver la meilleure manière d'exprimer les contraintes (sous forme strictement binaire, sur une échelle ordinale ou numérique), pour construire une hiérarchie de contraintes qui permette de rendre compte des réalisations observées, enfin pour déterminer la description la plus favorable de l'interaction de ces contraintes (stricte hiérarchie ou interaction plus complexe (Plénat 1996)).

En ce qui concerne spécifiquement la contrainte de grammaticalité, un embryon de réponse à ces questions est fourni par l'examen d'un sigle tel que RICM²⁸. Ce sigle, pour lequel seule l'épellation est attestée, ne viole aucune des contraintes mentionnées ci-dessus, à l'exception peut-être de la contrainte de grammaticalité. La séquence /km/ est pourtant attestée en tant que coda syllabique dans quelques formes, telles que *drachme*. Il est en conséquence difficile d'affirmer qu'elle est complètement imprononçable. Justifier le mode d'oralisation préférentiel de cette forme passe donc nécessairement par l'introduction d'une gradualité dans l'expression de la prononçabilité. Reconsidérée en ces termes, /aikm/ n'est probablement pas assez grammaticale, en comparaison avec la forme épelée, qui a pour elle de respecter la contrainte de longueur. Inversement, on peut concevoir que ces mêmes contraintes sélectionneraient la lecture d'un sigle (hypothétique) tel que STRICM²⁹, dont l'épellation est moins acceptable (violation de la contrainte de longueur).

Le niveau post-lexical

Pour le français, une nouvelle difficulté apparaît lorsque l'on envisage la transcription d'unités linguistiques de taille supérieure au mot : le problème

26. American Standard Code for Information Interchange

27. Il faudrait, en toute rigueur, utiliser des transcriptions plus abstraites, incluant en particulier les séparateurs entre lettres dans les formes épelées, de manière à rendre compte de la manière différentielle dont les contraintes pèsent sur chacun des types d'oralisation (cf. la note¹⁸).

28. Régiment d'Infanterie et de Chars de Marine

29. Service de Transport Routier Intercommunal par Camionnettes et Motocyclettes

de la liaison. Sans entrer dans une description complète du problème, pour laquelle nous renvoyons principalement à (Encrevé 1988), disons simplement que la prédiction correcte des liaisons renforce le besoin de disposer au moins d'un étiquetage morpho-syntaxique précis, voire d'une analyse syntaxique complète, au moment d'effectuer la transcription.

1.1.3 Les systèmes de transcription du français

Les approches par règles

Un certain nombre de systèmes de transcription automatique ont été développés, pour la quasi-totalité des langues européennes, pour répondre aux divers besoins que nous évoquions en tête de ce chapitre. En plus des systèmes déjà mentionnés, signalons pour le français, parmi les développements les plus récents, le système de transcription du LADL (Laporte 1988a), le système TOPH développé à l'ICP (Aubergé 1991) (dans lequel on trouvera des références historiques plus complètes), les divers systèmes VARION développés au LIMSI (Lacheret-Dujour 1990), le système GEPH développé à l'IRIT (Tihoni & Pérennou 1991), le système GRIPHON (Béchet, Derderian, & El-Bèze 1995), etc.

Tous ces systèmes partagent la caractéristique de fonctionner exclusivement sur la base de règles de réécriture dépendant du contexte³⁰ qui, étant donnés une forme graphique et un ensemble de traits morpho-syntaxiques associés, produisent une transcription par application en série ou en parallèle de règles du type $x \rightarrow y / C_g - C_d$, où x , y , C_g et C_d sont, suivant les systèmes et les étapes de traitement, des chaînes graphémiques ou phonémiques. Par exemple, dans notre système de transcription par règles (Yvon 1993), une chaîne graphique est réécrite en trois temps (à l'issue d'un traitement syntaxique réalisé par un analyseur par couches (Constant 1991)):

1. Normalisation orthographique généralisée. Cette étape du traitement, qui s'applique successivement à chacun des mots de l'entrée à transcrire, récrit des graphèmes en graphèmes, procédant, pour ainsi dire, à une vaste réforme de l'orthographe. Il existe plusieurs jeux de règles, qui ne s'appliquent qu'à des sous-ensembles des formes (règles spécifiques pour les verbes, pour l'ensemble du vocabulaire commun, pour les noms propres français, anglais, etc.). Ces jeux de règles s'appliquent en série et/ou en parallèle: en particulier, un nom propre est susceptible d'être récrit de plusieurs manières concurrentes selon la ou les provenances possibles que l'on aura pu identifier.

30. On considère ici qu'une entrée dictionnaire est une règle très spécifique, qui récrit intégralement un mot. On notera à cet égard que le rôle des dictionnaires dans les systèmes de transcription est en pleine mutation: originellement conçus pour stocker les formes les plus idiosyncrasiques, ils semblent, compte-tenu de la diminution importante des coûts de stockage et du développement de procédures spécifiques de gestion de gros lexiques semble aujourd'hui, devoir jouer un rôle plus sensiblement plus important. Dans ce nouveau schéma, c'est par accès lexical que sont directement transcrites la majorité des formes rencontrées, les règles n'étant plus utilisées que pour traiter les rares formes absentes du dictionnaire. En tout état de cause, les deux types de connaissances (règles générales et entrées dictionnaires) co-existent dans des proportions diverses dans les systèmes de transcription, et, sauf tarissement improbable de toutes les sources de création lexicale, devront continuer à co-exister.

À l'issue de cette étape, toutes les idiosyncrasies et irrégularités orthographiques ont été réduites. On notera que les sigles sont également concernés par cette réécriture.

Pour donner quelques exemples, *Fresnes* est récrit *Frêne*, *choriambe* est récrit *koriambe*, AP est récrit *âpé*, etc.;

2. Transcription graphème-phonème, par application d'un petit jeu de règles, qui correspondent approximativement à la connaissance de l'orthographe d'un enfant apprenant à lire. Ce traitement, comme le précédent, s'applique successivement et indépendamment à chacun des mots en entrée. Ce module produit des transcriptions très abstraites. Ainsi, pour continuer avec les mêmes exemples, à l'issue de ce traitement, *Fresne* est transcrit en /fɾɛnə/, *choriambe* en /kɔɾiambə/, AP en /apɛ/;
3. Applications d'une batterie de règles phonologiques (évaluation des E-muets, nasalisation des voyelles, assimilations diverses). Le domaine d'application de ces règles est la phrase toute entière; en particulier c'est durant cette étape que sont traités les problèmes de liaison. Cette dernière étape est susceptible de produire des transcriptions très variables (plus ou moins abstraites, plus ou moins réduites), suivant la batterie (paramétrable) de règles utilisée.

Dans l'ensemble, il existe pour les phénomènes linguistiques qui doivent être reproduits dans une telle chaîne de traitement une description linguistique plus ou moins complète (nous nous sommes appuyés principalement sur (Fouché 1959), (Warnant 1987) pour la conception des étapes de traitement (1) et (2), et sur (Dell 1985) pour l'étape (3)), et les dictionnaires électroniques aujourd'hui disponibles constituent un facteur supplémentaire d'accélération pour le développement de convertisseurs graphème-phonème.

Dans ce contexte, ce qui différencie principalement les systèmes de transcription, c'est principalement leur facilité de maintenance, c'est à dire leur capacité à être complétés, enrichis, chaque fois que nécessaire.

La majorité des systèmes développés tire en effet parti de la facilité qui consiste à ordonner les règles de la plus à la moins spécifique, de manière à simplifier la description du contexte des règles les plus générales. Dans un tel système de règles, la transcription s'effectue alors par une seule traversée, *sans retour arrière*, du jeu de règles; l'ensemble du traitement étant alors, malgré l'utilisation de règles dépendant du contexte, équivalent à une réécriture régulière, et implantable dans un automate d'état fini.

L'utilisation implicite de l'ordre des règles introduit malheureusement des dépendances complexes entre celles-ci, qui compromettent à terme, la maintenabilité de l'ensemble du système. D'où l'importance de disposer d'un langage de description des règles le plus complet possible, de disposer également d'outils de contrôle et de vérification de la cohérence de la base de règles.

Pour une approche à base de lexiques

Les approches à base de règles, en dépit de leur indéniable réussite, présentent un certain nombre de limitations. L'élaboration d'une base de règles fiable constitue un investissement lourd, investissement qui doit être renouvelé chaque fois qu'un système de transcription d'une nouvelle langue est développé. Au delà de cet argument économique, qui en lui seul justifie les approches fondées sur l'exploitation d'un lexique de transcriptions, nous voudrions avancer trois raisons supplémentaires, qui, en relation avec nos travaux sur les noms propres, nous incitent à plaider, du moins en ce qui concerne le traitement des mots isolés (correspondant aux étapes 1 et 2 de la chaîne de traitement de notre système de transcription), pour le développement d'approches alternatives, fondées sur l'exploitation de lexiques :

- de manière générale, il existe une tendance marquée à assigner aux lexiques un rôle de plus en plus important dans les systèmes de transcription, les règles proprement dit n'intervenant qu'en dernier recours, pour transcrire les mots inconnus. Dans ce contexte, la question se pose de conserver et de maintenir en parallèle les deux sources de connaissance que constituent dictionnaires et règles de écriture. En particulier, il nous paraît fondé de s'interroger sur la nécessité de conserver des systèmes de règles quand de vastes lexiques sont par ailleurs disponibles, qu'il nous semble profitable de mettre à contribution, y compris pour transcrire des mots qui n'y figurent pas;
- l'introduction de règles spécifiquement adaptées au traitement des patronymes et toponymes introduit inévitablement, en dépit des précautions qui peuvent être prises, une augmentation importante de l'ambiguïté des graphèmes, et de complexité des règles de transcription. Maintenir, dans ce contexte, un système de transcription cohérent est une véritable gageure. Dès lors que la transcription met en jeu des milliers de règles, dont l'interdépendance, *via* des conditions d'ordonnancement, est complexe, il nous semble que le principe d'une transcription par règles perd beaucoup de son opérationnalité, justifiant ainsi, par contrecoup, une approche à base de lexiques, dont la maintenance est sensiblement plus aisée;
- le traitement des patronymes et des toponymes requiert, par rapport au traitement des entrées du lexique commun, la connaissance d'une information supplémentaire : l'origine du nom propre. Déterminer cette origine est une question complexe, dans la mesure où, d'une part, la granularité correcte de cette information est extrêmement difficile à évaluer (combien d'origines différentes distinguer ? quelles normes de prononciation instaurer pour chacune des langues identifiées ?), et où d'autre part le problème de la détermination de la langue d'origine d'une chaîne donnée, n'a pas, à ce jour, reçu de solution entièrement satisfaisante. Il semble à l'inverse que des méthodes fondées sur l'utilisation de lexiques permettent doublement de s'affranchir de ce problème :
 - dans un système à base de lexiques, les transcriptions sont produites par ressemblance avec des transcriptions déjà connues. Dans un tel contexte, la question de fixer, pour chaque langue, une norme de

transcription compatible avec le niveau d'exposition moyen des locuteurs d'une communauté à la langue d'origine ne se pose pas : la norme de transcription est implicitement dans les données;

- le même argument permet d'affirmer qu'une « bonne » approche à base de lexiques devrait être à même d'inférer des ressemblances entre les formes qui partagent une origine commune, faisant ainsi disparaître le besoin de procéder à une phase préliminaire d'identification de la langue.
- la reformulation du traitement des sigles sous la forme d'un mécanisme d'évaluation de candidats vis-à-vis d'un système de contraintes symboliques ou numériques a mis en évidence la nécessité d'exprimer certaines de ces contraintes (la grammaticalité d'un groupe consonantique, l'adéquation d'une forme graphique à la forme phonétique correspondante) sur une échelle au moins ordinale. Pour ces deux contraintes, la description linguistique est aujourd'hui trop incomplète (voir toutefois (Plénat 1993; Dell 1995) sur la formulation de règles de bonne formation des syllabes), pour que l'on puisse construire ces échelles sur la base de règles.

Il semble donc qu'il faille, dans ce domaine également, se tourner vers des approches à base de corpus ou de lexiques, qui du reste ont déjà été utilisées dans le contexte de la description du système phonotactique du français ((Wioland 1985; Laporte 1992)).

1.1.4 Conclusion

Nous avons présenté dans cette section quelques unes des difficultés majeures que pose la tâche de transcription graphème-phonème, en insistant principalement sur le cas du français. Nous avons également souligné certaines des limites des approches à base de règles, qui constituent, toutefois, encore aujourd'hui, les réponses les mieux adaptées au traitement du lexique commun, à la fois en termes d'implantation informatique, et en termes de qualité de transcription.

Ces limitations, particulièrement flagrantes dans le domaine du traitement des unités extra-lexicales, plaident en faveur du développement d'approches alternatives. La principale alternative, qui consiste à développer des dictionnaires de grande taille (Schmidt *et al.* 1993), présente un inconvénient majeur : comme le note Christel Sorin (1991), à partir de données américaines, mais la situation est très semblable, voire pire, en France :

(...) si un dictionnaire de 2 000 noms propres couvre 50 % des noms pris au hasard dans un annuaire, il en faut 6 000 pour couvrir 60 % et l'on atteint une asymptote aux environs de 62 % pour la couverture des noms d'un annuaire à partir des noms d'un autre annuaire (...)

Ce qui semble requis, ce sont donc non-seulement des dictionnaires de grande taille, mais également des programmes capables d'en tirer le meilleur profit, c'est-à-dire capables de s'appuyer sur ces lexiques pour traiter les noms qui n'y figurent pas.

1.2 Les modèles cognitifs de la lecture

1.2.1 Introduction

On esquisse ici les grandes lignes des débats concernant les modèles cognitifs de la lecture³¹. Un grand nombre de données permettant de décrire de manière assez fine les processus cognitifs mis en œuvre lors de la lecture ont été amassées par les psychologues au cours des deux dernières décennies. La raison de cet intérêt tout particulier accordé à la lecture est simple : la tâche de lecture, avec la tâche de décision lexicale (qui consiste à décider si une chaîne donnée est ou non un mot du lexique), se prête particulièrement bien à des protocoles expérimentaux permettant d'étudier les mécanismes d'accès lexical, ainsi que les déterminants qui influencent ces mécanismes.

Avant de commencer cet exposé, une remarque s'impose : si ces modèles s'appuient de plus en plus sur une mise en œuvre informatique qui permet de tester plus facilement leurs performances, il faut noter que les tests eux-mêmes répondent à une logique différente de celle qui président à l'évaluation des systèmes de transcription pour la synthèse de la parole. Il est en effet plus important, pour l'évaluation des modèles cognitifs, d'être capable de simuler, même grossièrement, une large gamme de faits expérimentaux, que de reproduire au mieux les performances de l'humain en termes de capacité à lire des mots. Pour cette raison, et quand bien même des évaluations seraient fournies dans les différents articles auxquels nous nous référons, nous ne nous intéresserons pas dans cette section aux performances des différents systèmes implantés pour tester tel ou tel modèle.

Par ailleurs, on trouvera dans (Forster 1990) ou dans (Fayol 1992) des descriptions plus complètes de l'ensemble des différents modules utilisés dans les modèles du lexique mental : en particulier du module spécialisé dans la reconnaissance des lettres qui composent le stimulus graphique, et dont le résultat est un pré-requis de la tâche de lecture ; et dans (Sprenger-Charolles 1992) un aperçu des différentes étapes du développement chez l'enfant des capacités mises en jeu dans les tâches de lecture, ainsi qu'une discussion sur les spécificités dans ce cadre du français par rapport à l'anglais.

1.2.2 Le modèle de « double-voie »

Un premier modèle, proposé par (Forster & Chambers 1973), postule que l'humain fait simultanément appel, lors de la lecture d'un mot, à deux mécanismes bien distincts. Une chaîne de caractères à lire est tout d'abord recherchée dans un lexique stocké dans la mémoire. Si elle y est trouvée, alors sa prononciation est directement disponible, puisque la représentation phonologique du mot fait partie des informations lexicales mémorisées. Si la recherche échoue, alors la prononciation inconnue est produite par un module de trans-

31. Tout au long de cette section, « lecture » sera systématiquement utilisé pour décrire le processus qui consiste à « lire à voix haute » une chaîne de caractères, c'est à dire plus précisément à émettre une séquence acoustique en réponse à un stimulus visuel composé de symboles graphémiques.

cription graphème-phonème. Ce module, qui fonctionne en parallèle du module de recherche lexicale, met en œuvre des règles de transcription déterministes qui mettent en correspondance de petites unités lexicales (principalement des lettres, ou des petits groupes de lettres) avec un phonème. Deux compétences sont donc successivement mises à contribution par ce module : une compétence de segmentation, qui s'exprime par le repérage des unités pertinentes pour la prononciation (et notamment la reconnaissance de digraphes tels que *sh* ou *ea*); et une compétence de prononciation, qui associe un phonème à chacune des unités reconnues.

Ce modèle s'accorde bien avec la tradition dominante en linguistique à l'époque de sa publication, qui postule que les connaissances langagières sont effectivement apprises et représentées sous la forme de règles. Il introduit donc de façon naturelle les concepts de *mot régulier* et d'*exception* : les premiers peuvent être prononcés correctement par les deux méthodes, alors que la prononciation correcte des seconds dépend de leur présence dans le lexique. Il permet également de rendre compte d'un bon nombre de faits expérimentaux bien attestés : par exemple que les mots connus sont prononcés plus rapidement que des mots inconnus ou que des pseudo-mots³² : alors qu'il suffit pour prononcer un mot connu de retrouver son entrée lexicale, il faut pour un mot inconnu attendre que le module à base de règles, censément plus lent, achève son traitement.

Une autre caractéristique de ce modèle mérite une mention dans cette courte introduction. Pour rendre compte de l'influence expérimentalement attestée de la fréquence d'un mot sur la vitesse d'accès aux informations disponibles pour ce mot, Forster et Chambers font l'hypothèse que la recherche dans le lexique mental se fait par ordre de fréquence décroissante. Ceci entraîne en particulier que pour des mots rares, il est possible que le module à base de règles soit en mesure de proposer une prononciation avant que la recherche ne soit achevée, et que donc, certains mots peu fréquents soient prononcés uniquement sur la base de règles. Il est à noter que ce cas de figure n'est possible que pour des mots réguliers, c'est-à-dire dont la prononciation respecte les règles usuelles. Ceci est d'ailleurs conforme à des observations ultérieures, qui ont mis en évidence un désavantage des mots rares et irréguliers sur des mots rares et réguliers : alors les premiers ne peuvent s'en remettre qu'à la procédure de recherche pour la production d'une prononciation; les seconds compensent leur faible fréquence par leur régularité, et sont lus plus rapidement que les précédents par l'intermédiaire des règles de transcription.

L'existence de règles de transcription graphème-phonème est également supposée dans d'autres modèles (voir par exemple (Coltheart 1978)) qui postulent que l'accès lexical se fait en parallèle en utilisant simultanément des codes d'accès orthographiques (le stimulus tel qu'il est présenté) et phonologiques (la transcription de ce stimulus par la base de règles). Dans cette version, les règles de transcription jouent donc un rôle supplémentaire, puisqu'elles permettent également d'accélérer dans certains cas la recherche lexicale. Un modèle complet d'accès lexical intégrant ces diverses hypothèses est représenté à la figure 1.1.

Un dernier élément important mérite d'être souligné : ces modèles dits de

32. Un pseudo-mot est une chaîne alphabétique qui, bien que ne correspondant pas à une entrée lexicale, respecte la phonotactique de la langue. Ainsi, *bamin* est un pseudo-mot du français, à la différence, par exemple, de *dvriktp*.

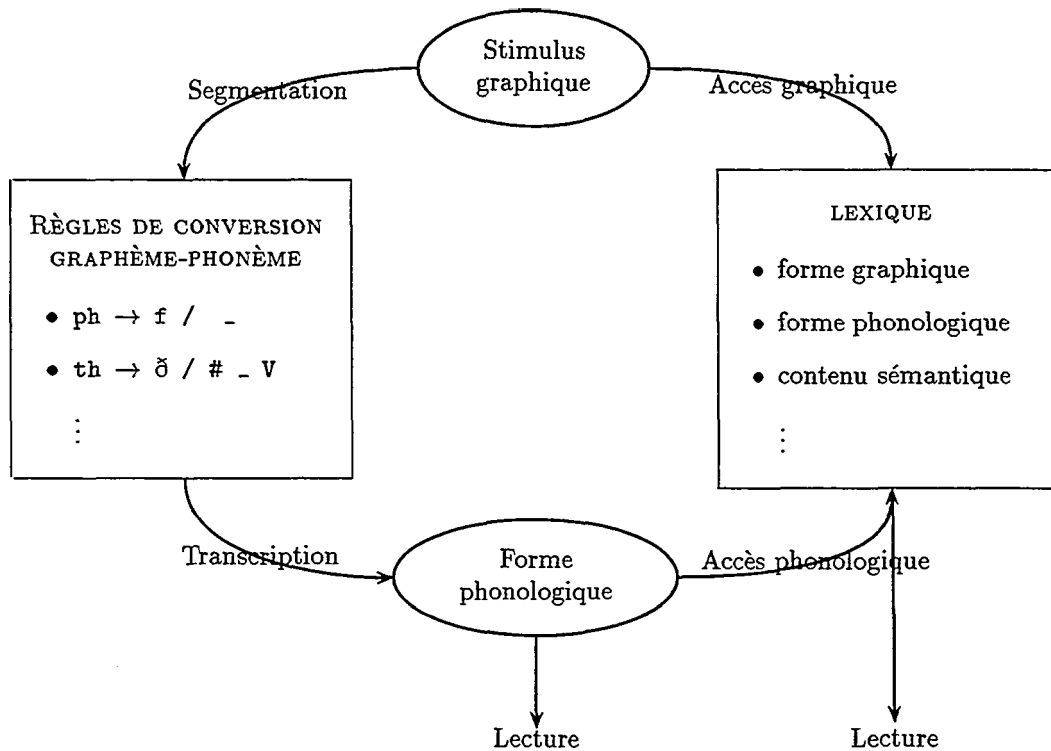


Figure 1.1: Un modèle de double accès lexical.

« double-voie d'accès » (*dual-route models*), présentent l'avantage de permettre de rendre compte de l'existence de certains troubles (dyslexie) de fonctionnement du langage, troubles qui se manifestent en particulier par des difficultés à exécuter correctement des tâches de lecture. Il existe en effet divers types de dyslexies (Casalis & Lecocq 1992), chacune correspondant à des symptômes très différents : ainsi les sujets atteints de *dyslexie profonde* sont incapables de prononcer des mots inconnus, ou des pseudo-mots, alors que la lecture de mots connus (réguliers ou irréguliers) ne pose pas de problème ; par contre les sujets atteints de *dyslexie de surface* sont incapables de prononcer correctement des mots irréguliers, mais lisent sans erreur des mots réguliers comme des pseudo-mots. Un modèle de double-route permet immédiatement de rendre compte de ces deux types de problèmes, en postulant que le premier type correspond à un dysfonctionnement du module de conversion graphème-phonème (mais la voie d'accès lexical reste fonctionnelle), et que le second correspond à un dysfonctionnement du module de recherche lexicale (mais la lecture des formes régulières reste possible, par le biais des règles de conversion). Cet argument, souvent invoqué pour justifier la pertinence de l'hypothèse des deux voies, gagnerait nous semble-t-il en force si lui était adjointe une réelle modélisation des dysfonctionnements qu'il postule.

1.2.3 Le rôle des analogies

Glushko (1979; 1981) critique ce modèle en remettant en cause l'existence de ces deux modalités (accès lexical et règles de conversion) très différentes de lecture. Son argument repose principalement sur la constatation expérimentale que le voisinage lexical (constitué par les mots qui présentent une forte ressemblance orthographique) d'un pseudo-mot semble conditionner le temps nécessaire à sa lecture, et influencer la construction de sa prononciation.

Soient ainsi les pseudo-mots (anglais) *taze* et *tave*, tous deux construits à partir de mots du lexique en substituant la première lettre. Glushko constate que le premier a deux voisins lexicaux, *daze* et *gaze*, dont la prononciation est régulière (i.e. suit les règles de prononciation de l'anglais); et qui de plus s'accordent sur la prononciation (/e/) du noyau vocalique. Le second a également plusieurs voisins lexicaux, tels que *have* et *gave*; mais si *gave* est régulier et se prononce /geiv/, *have* exhibe une prononciation alternative et irrégulière du noyau vocalique: /hæv/. Ce que Glushko observe alors, c'est que les pseudo-mots de type *taze*, i.e. dont l'environnement lexical est *cohérent* se prononcent plus rapidement que ceux du type *tave*, pour lesquels des prononciations irrégulières (i.e. non conformes aux règles de prononciation standard), manifestement influencées par la représentation phonologique de leur(s) voisin(s) sont proposées: ainsi la prononciation /tæv/ pour le mot *tave* lui-même, qui témoigne de l'influence du voisin *have*.

Il semble donc qu'il faille prendre en compte un nouveau déterminant influençant les temps de lecture: la cohérence de l'environnement lexical du mot, qui se manifeste à travers un effet « de gang » (*gang-effect*).

Or ces faits sont manifestement difficilement compatibles avec le modèle de Forster: si les deux types de pseudo-mots sont prononcés par des règles qui s'appliquent graphème par graphème, il n'y a pas de raison que les temps de lecture soient différents, ni que l'on observe des hésitations ou des incohérences concernant la manière de lire certains groupes de lettres. Pour sauver un modèle à base de règles, il faudrait postuler qu'elles opèrent sur des unités lexicales (rimes, pans de mots) qui sont bien supérieures à la lettre, pour pouvoir rendre compte des phénomènes observés. Chaque entrée lexicale étant elle-même une règle très spécifique, il devient inutile de faire l'hypothèse de deux mécanismes différenciés dont l'un traiterait les mots en tant qu'entités, et l'autre de larges portions de chaînes. Partant, l'hypothèse de deux voies d'accès au lexique devient également superflue. C'est la conclusion à laquelle parvient Glushko, qui fait l'hypothèse que la lecture des mots comme celle des pseudo-mots est le résultat d'un processus unique d'activation en parallèle des représentations lexicales de tous les mots *analogues* à la chaîne à prononcer.

Il est important de noter que l'analogie telle que l'introduit Glushko est assez floue: dans son esprit, l'analogie entre deux mots est conditionnée par le nombre de caractères qu'ils ont en commun, sans plus de détail. Les analogues des pseudo-mots qu'il donne à lire, et dont il suppose qu'ils sont mis à contribution sont précisément ceux qu'il a utilisés pour construire les pseudo-mots. Par construction, ils se trouvent donc tous à distance lexicographique unité des chaînes à prononcer. Toutefois, tous les mots à distance unité ne semblent pas

jouer le même rôle : ainsi l'existence des mots *tale* (/teil/), *tape* (/teip/) et *tare* (/tæər/) dans le voisinage de *tave* et *taze* ne semble jouer aucun rôle; il n'en est du moins pas fait mention. Implicitement, la rime semble donc être un élément privilégié du mécanisme d'analogie tel que ce modèle le postule. On trouvera dans (Taraban & McLelland 1987) et dans (Treiman & Zukowski 1988) des éléments expérimentaux complémentaires sur le rôle joué par la rime dans la construction d'analogies pour la lecture.

Plus qu'un modèle de prononciation par analogie, c'est donc un modèle de prononciation par activation-synthèse que propose Glushko : un stimulus graphique active en parallèle plusieurs entrées lexicales proches du stimulus, qui entrent en concurrence pour produire une synthèse acceptable de leur différentes prononciations. Ce modèle laisse néanmoins en suspens deux questions importantes : d'une part, il ne dit rien des mécanismes de recombinaison nécessaires à la synthèse des prononciations (dans le cas de *tave*, par exemple, d'où vient que le *t* se prononce /t/?); d'autre part, et ce point est intimement lié au précédent, ce modèle ne spécifie pas la façon dont les mots sont segmentés en unités lexicales de taille inférieure au mot.

On notera que cette hypothèse d'un mécanisme unique se trouve reprise dans les mises en œuvre connexionnistes des modèles de lecture (Seidenberg & McClelland 1989). Contrairement à Glushko, néanmoins, ces théories vont jusqu'à remettre en cause l'existence même d'un lexique mental, dans lequel les mots seraient représentés en tant que tels. Il semble pourtant (Coltheart *et al.* 1993) que la représentation distribuée du lexique sous la forme de patrons d'activation, représentation qui seule est postulée dans les modèles connexionnistes, rende compte de façon insatisfaisante d'un bon nombre de faits expérimentaux importants liés aux phénomènes de dyslexies. XXX De surcroît, malgré des améliorations récentes (Plaut & McClelland 1993), ces modèles présentent encore des capacités de généralisation insuffisantes par rapport à celles de l'humain.

Signalons enfin qu'une autre forme d'analogie, fondée sur les ressemblances phonétiques, semble également jouer un rôle dans les mécanismes de lecture. C'est du moins la conclusion à laquelle parviennent Brown et Besner (1986), à partir de l'observation du fait expérimental troublant suivant. Le groupe consonantique *cc* se lit, en anglais, /k/ ou /ks/, selon la nature de voyelle qui suit. Si c'est un *e* ou un *i*, la prononciation est *ks* (comme dans *accident*); sinon c'est /k/ (comme dans *accusative*). Cette règle semble ne souffrir pratiquement aucune exception. Pourtant, leurs expériences de lecture de pseudo-mots (par exemple *haccid*) révèlent une proportion anormalement élevée de lecture avec un /s/. Une explication possible est que cette prononciation est produite par analogie avec celle du mot *acid*. Cette première hypothèse, compatible, quoiqu'au prix d'une extension de la notion d'analogie, avec la théorie de Glushko, n'est pas confirmée par leurs autres expériences. Ce qui est mis en évidence, au contraire, c'est l'influence (l'effet d'attraction) de chaînes phonétiques correspondant à des mots *qui ne sont pas des analogues orthographiques*. Par exemple, malgré la régularité de la prononciation du segment *oun* en anglais (lu presque systématiquement /awn/), le pseudo-mot *joun* est fréquemment lu comme le mot *june*. Ces observations, qui s'accordent assez mal avec l'ensemble des modèles présentés ici, sont notamment à l'origine du système élaboré par Sullivan et Damper (1990; 1991). Retenons-en l'idée que la pression analogique des voisins lexicaux

semble s'exercer sous diverses formes, et de manière relativement variable, de manière d'autant moins prévisible que les mots à prononcer sont complexes (i.e. plus complexes que ceux de Glushko).

1.2.4 Éléments d'une synthèse

À la suite des découvertes de Glushko, et malgré la mise en doute de la portée de certains de ses résultats (Taraban & McLelland 1997), un certain nombre d'amendements ont été apportés au modèle de double-route de façon à prendre en compte ces faits nouveaux.

Ainsi, Rosson (1985) propose de réhabiliter le concept de règles de transcription, qui seraient associées à un module fonctionnant sur la base d'associations lexicales. Dans son modèle, toutefois, les règles sont ambiguës, c'est-à-dire qu'à un graphème (ou à une suite de graphèmes) peuvent correspondre plusieurs (suites de) phonèmes différents, parmi lesquels une sélection est faite sur des critères fréquentiels : une règle est d'autant plus forte (probable) qu'il y a de mots du lexique qui la satisfont. Sans rentrer dans le détail de ses expérimentations, qui mettent en évidence un effet de la force des règles sur les temps de lecture, notons qu'elle renvoie dos-à-dos et les partisans du modèle de double voie dans sa forme originelle, et les partisans d'un modèle d'une voie d'accès unique fonctionnant par activation lexicale. Les deux niveaux de connaissance, les connaissances alphabétiques liées à des unités infra-lexicales, et tout spécialement aux lettres; et les connaissances lexicales liées aux mots jouent un rôle manifeste dans les processus de prononciation. Tout modèle de la lecture doit alors intégrer ces deux types de connaissances, et proposer une modélisation plausible de leur interaction.

On trouvera finalement un exposé complet d'une synthèse plus récente et plus complète dans (Coltheart *et al.* 1993). Si effectivement le modèle de double-voie qu'il présente intègre partiellement les effets mis en évidence par Glushko, ainsi qu'un mécanisme d'apprentissage des règles de transcription, on peut penser que la manière dont ces effets sont reproduits laisse encore dans l'ombre un certain nombre de questions importantes. En particulier, le rôle particulier joué par la rime dans les phénomènes d'analogie est plus nié que reproduit...

En conclusion, ce tour d'horizon des différents modèles cognitifs de la lecture nous aura permis d'une part de présenter le modèle le plus plausible, celui du double-accès, ainsi que certains des faits expérimentaux importants qu'il permet d'expliquer. Il nous aura également permis de recenser une partie des compétences nécessaires à l'accomplissement de cette tâche : capacités de segmenter une chaîne en unités pertinentes (qui va sans doute au-delà de la capacité de segmenter en graphèmes, et inclue probablement des compétences en termes d'analyse morphologique (Caramazza, Laudanna, & Romani 1988)), capacité de transcrire ces unités pertinentes à l'aide d'un système complexe de règles, nécessité de disposer d'un lexique doublement accessible (par voie graphique et phonologique). Il nous aura enfin permis d'explicitier en quel sens les phénomènes d'analogie peuvent jouer un rôle dans ce domaine, même si, nous l'avons vu, la prise en compte de ces phénomènes dans un modèle de double-route est encore largement incomplète. En particulier, le rôle spécifique de la rime comme

la définition précise des éléments lexicaux capables d'influencer, sur la base de leur ressemblance graphique ou phonétique, la prononciation d'un mot, reste mal comprise et modélisée.

1.3 Quelques notes sur l'analogie

Nous initions, dans cette section, une discussion du concept d'*analogie*, en relation d'une part avec les modélisations proposées dans le domaine de l'intelligence artificielle, et d'autre part avec les développements effectués autour de ce concept dans le domaine du traitement du langage naturel.

Compte-tenu de la multiplicité des travaux dans ce domaine, notre ambition dans cette courte introduction sera nécessairement modeste : il s'agira avant tout de donner les quelques idées essentielles de la démarche analogique, et de signaler les réalisations avec lesquelles nous sommes le plus en phase. Nous donnerons tout d'abord une idée d'ensemble de la démarche consistant à tirer profit d'une base d'exemple pour apporter des solutions à de nouvelles situation, démarche exemplairement mise en œuvre dans les techniques de *raisonnement par cas*. Nous discuterons ensuite de la notion d'*analogie structurelle*, et essaierons de montrer pourquoi l'extraction d'analogies pertinentes est un problème difficile. Nous dirons ensuite un mot des quelques applications qui ont été faites de ces techniques dans le cadre du traitement du langage naturel.

1.3.1 Raisonnement par cas

Il est couramment, et de longue date, admis que la capacité à effectuer des analogies est une compétence cognitive essentielle et centrale chez l'humain. Cette compétence est supposée être mise à contribution dans un grand nombre de situations et pour résoudre des tâches variées de raisonnement (résolution de problème par identification d'un problème connu similaire, critique ou renforcement d'un argument sur la base d'exemples positifs ou négatifs), d'apprentissage (généralisation d'hypothèses), et de communication (compréhension de métaphores ou d'histoires, etc) (Haton *et al.* 1991, chap. 9). Cette compétence est volontiers associée à l'identification d'une réelle créativité de l'intelligence humaine. On retrouve une conception semblable de l'analogie au principe de nombreux tests d'« intelligence », qui comprennent souvent une tâche de complétion de schémas analogiques (Hoffman 1995). La compréhension et la modélisation informatique de cette compétence sont donc des thèmes de recherche populaires, correspondant à des objectifs importants, dans les domaines de l'intelligence artificielle et des sciences cognitives.

Le domaine particulier dans lequel la majorité des avancées récentes en la matière ont été menées est le raisonnement par cas (Kolodner 1993). Dans ce paradigme, deux grandes familles de tâches ont été identifiées pour lesquelles l'utilisation d'analogies se révèle particulièrement adaptée : la génération de critiques ou d'arguments, et la résolution de problèmes. D'une manière très générale, l'utilisation d'un mécanisme fondé sur l'analogie dans le cadre spécifique de la résolution de problèmes requiert la disponibilité d'une base de problèmes

résolus (qui joue le rôle d'un modèle de la mémoire), et comporte deux étapes principales :

- recherche dans la base d'exemples d'un problème partageant des caractéristiques voisines du problème en cours de résolution (problème *analogue*);
- adaptation de la solution connue au nouveau problème, par un transfert de connaissances. Durant cette étape, d'autres solutions peuvent être utilisées conjointement à la solution du problème analogue;

La mise en œuvre concrète de chacune de ces deux étapes pose des problèmes particulièrement ardues, qui concernent (nous renvoyons à (Kolodner 1993) pour un exposé très complet des problèmes posés, des solutions proposées, et de leur relation avec les modèles cognitifs de la mémoire et du raisonnement) :

- l'organisation informatique de la base de problèmes, à la fois au niveau de sa structure (pour optimiser la recherche), de son contenu (quel est le bon niveau de description des problèmes pour garantir à la fois qu'ils seront retrouvés à bon escient, et que leur réutilisation sera efficace), et du formalisme de représentation utilisé pour décrire ce contenu (logique, réseaux sémantiques, structures (*frames*)). Dans ce contexte se pose également la question de l'indexation, qui constitue une solution adaptée au problème de la recherche des problèmes potentiellement analogues, mais dont la conception (choix d'un vocabulaire, granularité de la description, dynamique de l'index) est souvent extrêmement délicate;
- la définition de la mesure de similarité entre problèmes, sur laquelle se fonde la recherche du problème analogue. Cette question est l'objet d'un enjeu particulièrement crucial sur le plan informatique, puisque l'on souhaite à la fois retrouver les situations qui partagent avec la nouvelle situation des ressemblances profondes (ou structurelles), et non pas seulement des analogies de surface, et en même temps avoir une fonction de calcul des similarités qui ne soit pas trop pénalisante au niveau des temps de recherche;
- les procédures de raisonnement mises en œuvre au niveau de l'étape de transfert de la solution connue. Cette étape requiert en effet que le système analogue soit à même de mettre en correspondance les sous-parties des problèmes qui jouent un rôle équivalent, et à effectuer un raisonnement pour décider ce qui doit être modifié, et quelle modification doit être opérée, de façon à adapter l'ancienne solution à la nouvelle situation.

1.3.2 Analogie structurelle et analogie de surface

Un des aspects réellement critique de ce type de système porte sur la phase d'identification des problèmes pertinents au regard de la nouvelle situation, identification qui repose sur un calcul de similarité. Or, il existe deux manières principales de concevoir la similarité entre problèmes, qui correspondent sensiblement aux deux acceptions principales du terme *analogie*. Ce terme possède en effet un sens « commun », dans lequel il est synonyme de *ressemblance* et de

similitude (sens (1)), et un sens plus « technique », plus respectueux de l'étymologie également, dans lequel l'analogie correspond à un rapport de proportion mettant en jeu quatre termes (sens (2)). Dans cette seconde acception, un objet (au sens large, ce peut être un système, un ensemble, un concept, une description, etc.) A sera dit l'analogue de B si on peut identifier deux autres objets C et D qui présentent entre eux un rapport de proportion identique à celui qui existe entre A et B , rapport de proportion qui se glose par « A est à B ce que C est à D ». Ce type d'analogie est donc toujours défini relativement à un contexte, qui est celui du rapport qui préexiste entre C et D , et en dehors duquel elle perd son sens. Une manière classique de représenter ce rapport de proportion est d'utiliser un graphique semblable à celui de la figure 1.2, sur lequel on a noté de R_1 la relation entre A et B , R_2 la relation entre C et D , et R_3 la relation associative par laquelle l'analogie s'établit. On utilisera également la notation $A : B :: C : D$ pour dénoter ce type d'analogies.

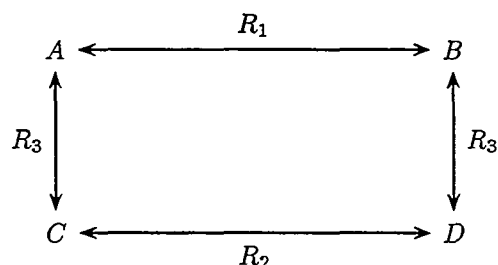


FIG. 1.2 - L'analogie comme un rapport de proportion

L'extraction d'analogies au sens (1) ne pose que peu de problèmes techniques, modulo la définition d'une fonction de similarité entre objets, exemplairement fondé sur l'évaluation des caractéristiques que deux objets partagent ou qui les opposent. Dans ce contexte, les diverses techniques connexionnistes peuvent également être considérées comme réalisant des formes particulières d'extraction d'analogie. L'extraction d'analogies au sens (2), qui requiert l'identification de rapports de proportion, et la mise en correspondance structurale ou fonctionnelle entre les objets A et C d'une part, et B et D d'autre part, est une tâche dont l'automatisation est notoirement plus complexe, et a donné lieu à de multiples tentatives de formalisation et de mise en œuvre (voir par exemple (Falkenhainer, Forbus, & Gentner 1990; Thagard *et al.* 1990; Mitchell & Hofstadter 1990) pour trois réalisations emblématiques en ce domaine, SME, ACME et Copycat). Si cette extraction est tout particulièrement difficile à réaliser informatiquement, on peut considérer, à la suite d'Hoffman 1995, que cela tient principalement à la très grande flexibilité, chez l'humain, des processus d'appariements analogiques. Hoffman donne un aperçu de cette flexibilité, en considérant la multiplicité des situations, correspondant chacune à des formes différentes d'analogies, que l'on obtient en faisant varier le sens et la nature des relations R_1 , R_2 et R_3 figurées sur le schéma 1.2.

La situation la plus simple est celle où R_1 est identique à R_2 (qu'il s'agisse d'une relation marquant une forme d'inclusion ou traduisant une dépendance fonctionnelle) et où R_3 traduit juste l'identité des relations. L'analogie entre le soleil et la terre d'une part, et l'électron et le noyau atomique d'autre part, est

un exemple de cette situation. La terre est en effet en rotation autour du soleil, de même que l'électron est en rotation autour du noyau : $R_1 (= R_2)$ exprime cette double relation qui unit un satellite à son attracteur. L'analogie est établie par simple identification (par R_3) des forces qui provoquent ce mouvement de rotation. Dans ce cas également, le sens de l'analogie est indifférencié, elle fonctionne à l'identique dans les deux directions, et toute analogie $A : B :: C : D$ peut aussi s'écrire $A : C :: B : D$ (la terre est à l'électron ce que le soleil est au noyau atomique).

En s'appuyant sur des exemples dans le domaine des analogies verbales-conceptuelles, Hoffman identifie, à côté de ces situations simples, une large palette de situations sensiblement plus complexes, dans lesquelles par exemple R_1 et R_2 sont différentes, ou dans lesquelles R_3 n'est pas à proprement parler inversible, ou pour lesquelles l'analogie joue simultanément sur plusieurs dimensions sémantiques, ou encore pour lesquelles le niveau d'abstraction des relations entre objets est élevé. Sans entrer dans un listage exhaustif du florilège dressé par Hoffman, mentionnons simplement la suivante, qui donne une bonne idée de la complexité que peuvent atteindre les analogies dans ce domaine : *ThisAnalogy:Analogy::"Tautalogie":."Tautology"*.

Il semble donc exister deux visions relativement distinctes des processus analogiques, l'une mettant en jeu des calculs de similarité, et l'autre des mises en correspondance structurelles et/ou fonctionnelles plus abstraites. Il serait pourtant réducteur d'appréhender cette dichotomie sous la forme d'une opposition tranchée entre une version « simpliste » et une version plus « réaliste » des phénomènes d'associations analogiques; le rapport entre ces deux mécanismes devant plutôt s'envisager sur le mode de la complémentarité, tant sur le plan cognitif que sur le plan informatique (Forbus, Gentner, & Law 1994). Il semble en effet que la capacité d'un sujet humain à extraire des analogies structurelles dans un domaine soit intimement liée à sa connaissance de ce domaine. Il semble également, du point de vue informatique, qu'il soit nécessaire de faire conjointement appel aux deux types de mécanismes liés à ces deux conceptions : le calcul d'analogies au sens (1), moins coûteux algorithmiquement, servant alors de filtrage de la base d'exemples, réduisant ainsi le champ des objets susceptibles de participer à une analogie de type (2).

1.3.3 Analogie et traitement du langage naturel

En ce qui concerne l'utilisation de techniques à base d'analogies dans le domaine du traitement du langage naturel, il est possible d'opérer grossièrement le même type de dichotomie, entre les approches qui privilégient le calcul et l'utilisation de similarités, et celles qui se focalisent sur l'identification de rapport de proportions.

Considérées sous l'angle d'un calcul de similarité, les techniques à base d'analogies sont susceptibles d'extraire des patrons de régularités, et de fonder des systèmes capables de simuler un comportement à base de règles, avec un avantage supplémentaire, celui de la flexibilité (Skousen 1989). À ce titre, de tels systèmes constituent des alternatives opératoires à tous les mécanismes de traitement du langage naturel qui reposent sur la manipulation de règles d'experts.

D'innombrables réalisations ont été accomplies dans cette direction de recherche, concernant la désambiguïsation sémantique (identification du « sous-sens » correct) dans des tâches de traduction ou de compréhension de texte, concernant la sous-catégorisation de verbes, la résolution d'ambiguïtés syntaxiques (rattachement des groupes prépositionnels, etc.), l'analyse morphologique, la transcription graphème-phonème, etc. Nous n'entrerons pas ici dans l'énumération exhaustive de ces différentes approches, et renvoyons aux actes des conférences et séminaires explicitement consacrés aux approches à base de données pour le traitement du langage naturel, et particulièrement à (Powers & Reeker 1991; Daelemans & Powers 1992; Adriaans 1993; Jones 1994; Wermter & Riloff 1995). Une présentation détaillée des approches à base de données pour la transcription graphème-phonème sera fournie au chapitre 2.

L'autre point de convergence entre raisonnement analogique et traitement du langage naturel concerne le problème de la compréhension des métaphores. La capacité humaine à produire et à comprendre des métaphores est en effet intimement liée à celle d'effectuer des analogies (Lakoff & Johnson 1985; Indurkha 1992), et tout spécialement des analogies au sens (2). Les développements dans ce domaine en sont pour l'instant, compte-tenu de la complexité de la modélisation, très limités, l'essentiel des contributions dans ce domaine se situant au niveau de la description.

1.4 Conclusion

Nous avons présenté dans cette section les éléments de contexte indispensables à la compréhension de notre étude. Nous avons en particulier souligné les limites des approches à base de règles pour la transcription graphème-phonème, en particulier en ce qui concerne les éléments extra-lexicaux. Nous avons également rappelé les principaux arguments avancés, dans le cadre de la psychologie cognitive de la lecture, en faveur d'un traitement analogique des mots inconnus. Nous avons enfin décrit les deux conceptions complémentaires des phénomènes analogiques telles qu'elles sont modélisées dans le domaine de l'intelligence artificielle, et du traitement des langues naturelles.

Pour conclure ce chapitre, mentionnons, en relation avec les deux derniers éléments de notre présentation, que les conclusions que Glushko tire de ses diverses expériences restent relativement évasives quant à la nature des analogies qui sont mises en jeu. Compte-tenu du vocabulaire employé, il est licite de supposer qu'il n'envisage que des analogies de surfaces (i.e. fondées sur la similarité entre les chaînes graphiques). Nous aurons l'occasion de revenir largement sur ce point, et de montrer que, compte-tenu de la forte structuration des lexiques dans lesquels les relations de similarité sont susceptibles d'intégrer d'autres éléments, tels que la ressemblance morphologique, et le partage de traits syntaxiques ou sémantiques, l'extraction d'analogies plus complexes peu se révéler à la fois fondée et fructueuse.

Chapitre 2

Prononcer/Classer

2.1 Préambule

Ce chapitre est consacré à un large tour d'horizon de différentes techniques qui ont été appliquées dans le passé au problème de l'apprentissage automatique des « lois » de conversion de l'écrit vers l'oral. Au delà des différences qui peuvent exister entre ces méthodes, et qui reflètent la grande variété des motivations (applicatives, démonstratives, etc.) ayant présidé à leur mise en œuvre, au delà également de la diversité des cercles académiques devant lesquels elles ont été présentées, ces méthodes partagent un certain nombre de caractéristiques importantes, qui justifient à nos yeux leur regroupement.

Hormis la reprogrammation d'un algorithme classificatoire prototypique (voir l'annexe B), notre contribution personnelle dans ce chapitre se situe essentiellement au niveau de la mise en forme d'un certain nombre de remarques et critiques, qui sont exposées dans la dernière section de ce chapitre. Pour le reste, il s'agira essentiellement d'une présentation de méthodes proposées par d'autres auteurs.

2.1.1 Les approches classificatoires

La principale des caractéristiques que partagent les approches ici présentées est qu'elles envisagent et posent le problème de la transcription graphème-phonème essentiellement comme un problème de *classification*¹. Dans tous les cas, la question de la segmentation de la chaîne orthographique, dont nous avons mentionné qu'elle constitue une compétence essentielle pour l'apprentissage de la lecture, est supposée résolue : un mot est composé d'une séquence de graphèmes atomiques, qui seront analysés et traités l'un après l'autre. En d'autres termes, si l'on se réfère aux théories cognitives exposées à la section 1.2, les compétences

1. Problème de classification peut s'entendre de deux manières différentes : comme le problème de la construction d'une partition d'un ensemble ; ou comme le problème de l'assignation, une partition étant donnée, d'une étiquette de classe à un élément dont l'étiquette est inconnue (problème de *discrimination* pour les statisticiens). C'est dans cette deuxième acception que nous utilisons ce terme..

en matière de décomposition d'une chaîne graphique en unités pertinentes sont modélisées ici par l'appariement terme à terme des symboles graphémique et phonémique. Transcrire un mot se ramène alors au problème d'assigner successivement à chacun des graphèmes le phonème (la classe) qui lui est associée dans la chaîne phonétique. On notera que cette démarche se distingue de ce point de vue des systèmes de transcription à base de règles, où il est courant de transcrire plusieurs lettres à la fois (par exemple *ph* en français).

Une seconde caractéristique importante de ces techniques d'apprentissage est qu'elles sont *supervisées*. Avant de pouvoir traiter correctement des mots nouveaux (non-appris), le système doit être nourri d'un lexique de mots dont la transcription est supposée connue non seulement dans sa globalité, mais aussi dans le détail précis de la mise en correspondance bi-univoque (ci-après *alignement*) des symboles des chaînes graphiques et phonétiques. Dans la mesure où les deux chaînes n'ont pas toujours le même nombre de symboles, au détriment le plus souvent de la chaîne phonétique (dans les langues que nous avons étudiées), cet alignement consiste essentiellement à assigner, du mieux possible, certains graphèmes à un « phonème » vide². Différents alignements du mot *chanson* avec sa transcription phonétique sont reproduits à la figure 2.1 (les phonèmes nuls sont représentés par /-/). Ces alignements mettent en évidence l'arbitraire de certaines correspondances : pourquoi préférer l'alignement d'une voyelle nasale avec la voyelle graphique (cas (a)) plutôt qu'avec la nasale graphique (cas (c)) ?

c	h	a	n	s	o	n
f	-	ã	-	s	õ	-

(2.1.a)

c	h	a	n	s	o	n
f	ã	s	õ	-	-	-

(2.1.b)

c	h	a	n	s	o	n
f	-	-	ã	s	-	õ

(2.1.c)

FIG. 2.1 - Quelques alignements de *chanson* avec /fãsõ/.

2.1.2 À propos de l'alignement

Les difficultés que soulèvent la construction automatique de telles représentations sont abordées en détail dans l'annexe A. Contentons-nous d'en souligner une, qui apparaît clairement lorsque l'on examine de près des mots tels que *voyage*, dont la représentation phonétique (/vwajaz/) se laisse difficilement aligner avec son pendant orthographique d'une façon qui satisfasse l'intuition. Une première solution est d'introduire de nouveaux symboles phonétiques pour se ramener à un alignement plausible : dans ce cas, introduire un symbole pour représenter le diphone /wa/ semble, à plus d'un titre, justifiable. Nous avons pour notre part préféré conserver la symétrie entre les deux représentations, et autoriser l'insertion de symboles « vides » dans la chaîne graphique. Les formes alignées sont alors (*voy-age*, /vwajaz-/). Ceci pose un problème manifeste au

2. On trouvera dans (Besling 1995) une interprétation légèrement différente de l'alignement, dans laquelle au lieu d'utiliser un symbole « nul », les chaînes sont littéralement « étirées » (par réplification de symboles) de façon à atteindre des tailles égales. L'impact sur les performances du système de classification de l'une ou l'autre conception est toutefois peu clair.

moment de la transcription de mots inconnus, dans lesquels, naturellement, ce « nul » graphique n'est pas représenté : il faut en conséquence classer des graphèmes qui ne figurent pas dans la forme à transcrire. Si la prise en compte de ce phénomène n'a pas été sans augmenter sensiblement la complexité de nos programmes³, au moins aura-t-elle permis qu'à bon compte, nos algorithmes de conversion graphème-phonème soient aussi capables du traitement inverse, celui de la transcription phonème-graphème.

2.1.3 Vue d'ensemble

Ce chapitre, destiné à dresser un état de l'art dans le domaine de l'apprentissage de la transcription graphème-phonème, sera également pour nous l'occasion d'en présenter les difficultés. Les descriptions linguistiques de cette correspondance ont largement mis en évidence l'intrication complexe⁴ de règles et d'exceptions qui la caractérise, ainsi que la nécessité (plus ou moins impérieuse, selon les langues) de construire ou d'identifier des unités graphémiques de niveau supérieur (syllabe, morphème) pour parvenir à un niveau de description satisfaisant. « Apprendre » par des moyens automatiques à réaliser cette transcription sans prendre en compte ces unités intermédiaires représente alors une entreprise délicate, comme en témoignent les performances encore insuffisantes de la grande majorité des systèmes développés à ce jour.

Ce chapitre sera finalement l'occasion de dresser un bilan en forme de critique des approches *classificatoires*, terme sous lequel nous désignerons les approches partageant les deux caractéristiques énoncées ci-dessus : transcription par classification successive des graphèmes, apprentissage supervisé de la classification sur des bases fréquentielles.

L'organisation de notre présentation rend compte de l'identification au sein de ces techniques de deux grands sous-groupes, suivant qu'elles se fondent sur une étape préalable d'apprentissage (section 2.2); ou qu'elles travaillent directement sur les données disponibles (section 2.3). Cette dichotomie rejoint en fait dans une large mesure la distinction classique en statistique entre techniques *paramétriques* (qui requièrent que les *paramètres* d'un modèle soient connus ou estimés) et les techniques non-paramétriques. La section finale élabore une synthèse critique de l'ensemble de ces expériences, en s'appuyant en particulier sur quelques résultats obtenus en mettant en œuvre l'algorithme DEC (Torkolla 1993). Cet algorithme, que nous avons été conduit à redévelopper intégralement, ainsi que l'ensemble des évaluations que nous avons menées, est décrit dans l'annexe B.

Avant de commencer ce tour d'horizon, notons finalement que les techniques mises en œuvre pour faire de la transcription graphème-phonème sont toutes susceptibles, *mutatis mutandis*, de s'appliquer à trois autres sous-tâches très

3. Tout au long de notre exposé, les algorithmes que nous avons développés seront présentés sous leur forme la plus simple, i.e. en faisant l'hypothèse qu'il n'y a pas de « nuls » dans la chaîne graphique. Néanmoins, leur extension au cas où cette condition n'est pas respectée sera suggérée à chaque fois que cela nous semblera nécessaire.

4. Le degré de complexité de cette correspondance est très variable selon les langues étudiées. Nous aurons l'occasion de le vérifier sur divers lexiques.

voisines :

- la conversion phonème-graphème. Le changement de perspective s'effectue en intervertissant les entrées et des sorties du système (voir par exemple (Rentzepopoulos & Kokkinakis 1991) pour un traitement probabiliste de ce problème);
- la détermination du ou des graphèmes qui porteront le ou les accents lexicaux. Cette identification est dans certaines langues un pré-traitement indispensable pour la transcription. Le changement de perspective se fait en remplaçant l'alphabet phonétique en sortie du système par un alphabet décrivant les positions accentuées (par exemple 0 pour une position non-accentuée, 1 pour une position accentuée). Concernant ce problème, on pourra se reporter à (Church 1985) pour une approche probabiliste; à (Daelemans *et al.* 1993; Gillis, Daelemans, & Durieux 1994) pour un traitement dans le cadre du paradigme « instance-based-learning »; à (Durieux 1992) ou encore à (Pirrelli & Salza 1993) pour la mise en œuvre et l'évaluation de techniques à base d'analogies;
- la détermination d'une segmentation en syllabes légales d'une chaîne de graphèmes ou de phonèmes. Le changement de perspective se fait en remplaçant l'alphabet phonétique en sortie du système par un alphabet permettant d'identifier les débuts ou fins de syllabe. On trouvera dans (van den Bosch & Daelemans 1992) une application à la syllabation de techniques connexionnistes.

2.2 Apprentissage de la transcription graphème-phonème

Dans cette section, nous présentons en détail l'application de diverses stratégies d'apprentissage⁵ pour la transcription graphème-phonème : une architecture connexionniste (section 2.2.1), une approche à base d'arbres de décision (section 2.2.2), et une approche probabiliste (section 2.2.3). Notre décision de regrouper dans une même section ces approches relativement différentes se fonde sur une similitude de surface : ce sont toutes des techniques qui n'utilisent pas directement les exemples disponibles pour transcrire des mots inconnus.

En effet, elles requièrent dans un premier temps que l'information pertinente contenue dans les données soit automatiquement extraite et intégrée (par exemple sous la forme de valeur de paramètres) dans des systèmes de classification dont la forme est figée. Ces systèmes de classification, une fois appris, serviront à traiter, *de manière uniforme*, les mots inconnus.

5. D'une manière ou d'une autre, tous les systèmes dont il sera question dans ce document peuvent se voir attribuer l'étiquette *apprentissage*. Nous utilisons ici ce terme dans un sens relativement limité : il y a apprentissage dans la mesure où l'exploitation des exemples pré-suppose une étape préliminaire de traitement, conduisant sous une forme ou une autre à la mise sous forme exploitable pour le mécanisme de transcription des données d'apprentissage.

2.2.1 Transcription par des méthodes connexionnistes

L'application de techniques connexionnistes au problème de la conversion graphème-phonème dans le système Nettetalk (Sejnowski & Rosenberg 1987) est régulièrement présentée comme une des réalisations les plus spectaculaires et les plus réussies de ces méthodes. C'est essentiellement cette approche, qui a été également appliquée au français (Laporte 1988b), que nous décrirons ici; signalons toutefois pour mémoire les travaux de (Lucas & Damper 1992) qui décrivent la mise en œuvre de réseaux connexionnistes syntaxiques; ainsi que ceux de (Hensen 94) qui rapportent l'application à la transcription graphème-phonèmes de réseaux inspirés des cartes de Kohonen (Kohonen 1984).

L'architecture de Nettetalk

Décrivons brièvement l'architecture utilisée par Sejnowski et Rosenberg (voir également figure 2.2). Il s'agit d'un classifieur fondé sur un perceptron multicouches, qui permet de classer un à un les graphèmes qui se présentent en entrée dans une classe unique correspondant au phonème à émettre. Pour tenir compte des effets de contexte, le réseau fonde sa décision sur une fenêtre graphémique de taille fixe (7) centrée sur la lettre à transcrire : on peut d'ailleurs aussi bien considérer que le système classifie en fait des fenêtres graphémiques de taille fixe. Chaque lettre étant codée sous la forme d'un vecteur de 29 bits (les 26 lettres de l'alphabet et trois symboles additionnels pour représenter la ponctuation et les frontières de mot), dont seul un est non-nul, le vecteur d'entrée est donc composé de $7 \times 29 = 203$ bits.

Les phonèmes sont codés en utilisant un système de traits distinctifs proche de celui couramment utilisé en phonétique. Le système de codage de Sejnowski et Rosenberg comprend 26 traits distinctifs, parmi lesquels cinq encodent des informations supra-segmentales (niveau d'accent pour les voyelles et diphtongues, et position dans la syllabe pour les consonnes). On notera que ce système de codage est particulièrement peu économique, puisqu'il permet de représenter 2^{26} symboles différents, alors que seules quelques centaines de combinaisons (phonème, information supra-segmentale) correspondent à des unités linguistiques effectives (ainsi toutes les combinaisons de consonnes et d'accents ne correspondent à rien dans la description de l'anglais). De plus, dans la mesure où la configuration de sortie du réseau ne correspond pas forcément à un symbole de l'alphabet phonétique utilisé, les sorties du réseau subissent un traitement additionnel, qui consiste à identifier le vecteur de sortie avec le phonème le plus ressemblant⁶.

Entre les 203 unités de la couche d'entrée et les 26 de la couche de sortie, Nettetalk intègre une couche cachée de 80 unités, qui sont complètement connectées aux unités d'entrée et de sortie. Au total, ce sont donc $(203 + 29) \times 80 = 18560$ paramètres (les poids associés aux différentes connexions) qui sont estimés durant l'apprentissage. Cet apprentissage est fondé sur une procédure itérative de rétro-propagation des erreurs.

6. Sejnowski et Rosenberg utilisent le produit scalaire entre le vecteur de sortie et le vecteur de bits représentant un phonème comme mesure de la similarité.

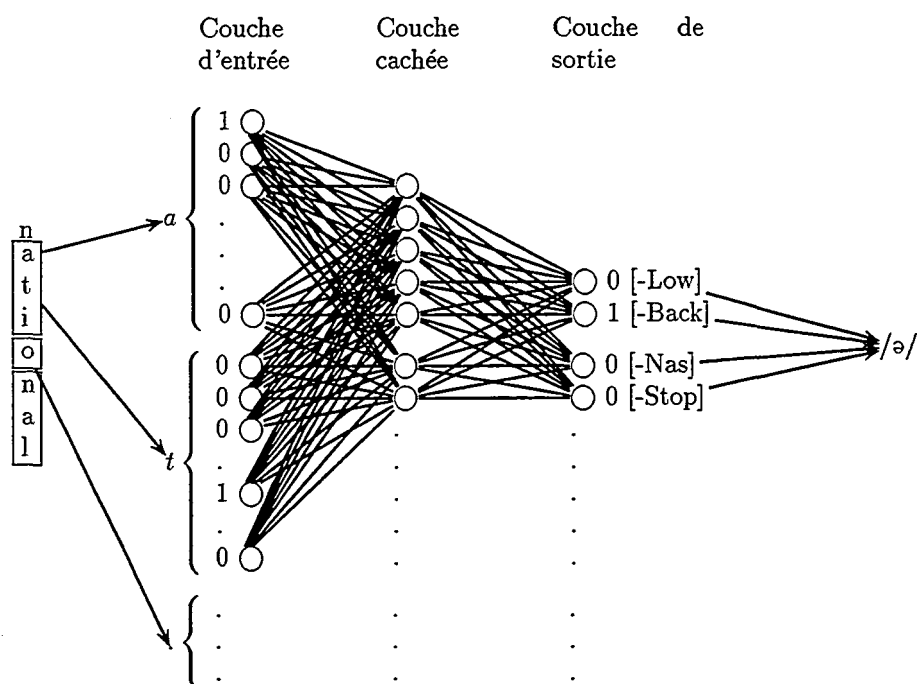


FIG. 2.2 - Architecture de Nettalk. Le graphème à classer est le o de national, le phonème correspondant est un /ə/.

Les résultats publiés par Sejnowski et Rosenberg permettent assez mal de se faire une idée précise des performances réelles que l'on peut espérer d'une telle architecture si l'on dispose d'un nombre suffisamment important d'exemples⁷. À partir des chiffres qu'ils rapportent, et de diverses autres mises en œuvre d'architectures similaires (McCulloch, Bedworth, & Bridle 1987; Ainsworth & Pell 1989; Dietterich, Hild, & Bakiri 1995), on peut estimer que le taux de phonèmes⁸ correctement prédits sur un jeu-test de mots non utilisés pour l'apprentissage est de l'ordre de 90-92%, ce dernier chiffre étant probablement une estimation assez optimiste des performances. L'évaluation à partir de tests d'écoute telle que celle mise en œuvre par Golding (1991) donne toutefois des résultats plus satisfaisants : dans son protocole, les performances du réseau connexionniste avoisinent les 67 % de mots corrects. On peut raisonnablement penser que si

7. Sejnowski aurait néanmoins, en utilisant une fenêtre graphémique de 11 caractères, et un lexique d'apprentissage de 18 000 mots, obtenu des résultats (non-publiés) de l'ordre de 92% de phonèmes corrects sur un lexique de test de 2 000 mots.

8. Diverses mesures des performances des systèmes de transcription ont été proposées : pourcentage de lettres correctement transcrites (classées), de phonèmes correctement prédits, de mots intégralement corrects. Si dans l'approche classificatoire, la première mesure se justifie, puisqu'elle correspond à l'évaluation la plus appropriée du système, elle présente toutefois l'inconvénient d'être inadaptée pour évaluer des systèmes de transcription pour la synthèse, où seuls les résultats en termes de phonèmes, voire de mots corrects sont réellement pertinents. Nous nous baserons en conséquence tout au long de ce document sur les deux dernières mesures ; on se reportera également à l'annexe D pour une (trop courte) discussion des problèmes d'évaluation.

Nettalk se trompe, ses erreurs sont pour une large part suffisamment minimales pour ne pas trop nuire à l'intelligibilité.

Discussion

Le modèle connexionniste mis en œuvre par Sejnowski et Rosenberg est une démonstration réussie de la capacité d'un réseau à induire, à partir d'un ensemble relativement petit de données, des règles de classification extrêmement complexes. Deux critiques peuvent néanmoins être adressées à ce modèle.

La première critique concerne les performances: si les résultats en termes de phonèmes correctement classifiés sont satisfaisants, il faut garder en mémoire le fait que ce qui importe pour la synthèse de la parole, ce sont les performances en termes de *mots intégralement corrects*. Les seuls résultats exprimés dans ces termes sont donnés dans (Dietterich, Hild, & Bakiri 1995), et confirment l'extrapolation faite par Coker (1990): avec des performances en termes de phonèmes corrects voisinant les 90 %, le pourcentage de mots transcrits sans erreur est de l'ordre de 50 à 60 %, soit bien en deçà des seuils de qualité requis dans les systèmes de synthèse de parole industriels.

La seconde réserve porte sur le type de représentation des phonèmes utilisée. Le choix d'une représentation de sortie distribuée plutôt que monolithique (une unité de sortie par phonème possible) introduit deux biais difficilement mesurables. Il est d'abord difficile d'isoler la part des performances qui doit être explicitement attribuée à la méthode d'identification finale entre vecteur de sortie et phonèmes. La seule indication est donnée par (McCulloch, Bedworth, & Bridle 1987), qui rapportent que dans le meilleur des cas, environ 40 % des vecteurs de sortie ne correspondent pas exactement à un phonème attesté. Le bon ou mauvais classement de ces phonèmes dépend alors uniquement de la procédure de post-classification adoptée. On aura l'occasion de revenir sur ce point dans la section 2.2.2.

L'utilisation d'une représentation distribuée des sorties introduit un second type de biais, dans la mesure où ce système de représentation rend le réseau capable d'exprimer des généralisations dont sont incapables d'autres systèmes de classification, qui se fondent sur une représentation monolithique des phonèmes. Par exemple, la proximité entre les phonèmes /p/ et /b/, qui ne diffèrent que par le trait de voisement, fait partie des connaissances fournies *a priori* et dont le système peut tirer parti. De même, le système est doté de la connaissance de toutes les combinaisons possibles phonème/accrntuation, y compris de celles qui *n'apparaissent pas* dans le jeu d'apprentissage: la procédure finale de réaffectation projette en effet le vecteur de sortie sur l'ensemble de toutes les combinaisons (phonème, marqueur supra-segmental) licites⁹. Si cet aspect particulier du traitement peut aussi bien être vu comme extrêmement positif, il n'en demeure pas moins qu'il introduit nécessairement un biais, dont il faut tenir compte dans les comparaisons avec d'autres systèmes qui utilisent une représentation plus pauvre.

Une dernière remarque sur ce système est d'ordre plus générale, et vaut de

9. On trouvera dans (Dietterich, Hild, & Bakiri 1990) des éléments permettant de penser que ce biais a une influence faible sur les résultats

fait pour tous les systèmes de transcription fondés sur des techniques de classification. Nous ne ferons donc que l'esquisser ici, en repoussant en section 2.4 de plus amples développements. Elle porte principalement sur les possibilités d'améliorer les performances pour les ramener à un niveau plus proche de celles de l'humain. Ni les expériences de McCulloch et al., ni celles de Ainsworth et Bell n'offrent à cet égard de perspectives encourageantes : ni l'augmentation sensible de la taille des ensembles d'apprentissage, ni l'augmentation de la taille de la couche cachée, ni l'augmentation de la taille de la fenêtre graphémique, ni même l'ajout d'une couche cachée supplémentaire ne permet de dépasser significativement la barre des 90 % de phonèmes corrects lorsque l'on fait prononcer des mots inconnus. Il semble qu'on bute sur une des limites intrinsèques de l'approche, sur laquelle nous reviendrons.

2.2.2 Transcription par induction d'arbres de décision

Popularisées par les travaux de Quinlan, en particulier par l'algorithme ID3, ((Quinlan 1983; Quinlan 1993)), les techniques de construction d'arbres de décision à partir de représentations symboliques ont été appliquées, sous diverses variantes, au problème de la transcription graphème-phonème. Nous présentons ici les travaux les plus convaincants à nos yeux, ceux de Dietterich et Bakiri (1990; 1991; 1993; 1995); on se reportera également avec profit à (Lucassen & Mercer 1984) pour une approche théoriquement fondée de la sélection des attributs utilisés pour classer un graphème; à (van den Bosch & Daelemans 1993) ou (Andersen & Dalsgaard 1993) pour l'exposition d'une mise en œuvre simplifiée de l'induction d'arbres de décisions. L'algorithme qui servira de référence pour nos évaluations, DEC (Torkolla 1993), et que nous présentons en détail dans l'annexe B, peut également être rattaché à cette famille de méthodes.

Règles et arbres de décisions

Avant d'entrer dans le vif du sujet, notons qu'une profonde affinité unit les techniques de construction d'arbres de décision et les méthodes d'induction de règles, les deux méthodes étant de fait équivalentes (Quinlan 1987). C'est la raison pour laquelle nous n'avons pas jugé nécessaire de consacrer une section indépendante aux techniques visant explicitement à l'extraction de règles. Notons simplement que ce qui distingue principalement ces deux approches, c'est la forme sous laquelle les règles sont extraites. Ainsi, un algorithme tel que ID3 construit des règles dont la forme est rigide et pré-définie : ce sont des conjonctions de prédicats portant sur les valeurs d'un attribut. Ces règles sont extraites par optimisation d'un critère numérique d'entropie. Ces deux caractéristiques contribuent à rendre les règles induites difficilement interprétables. Ce caractère est particulièrement exacerbé dans le cas de Dietterich et Bakiri, qui utilisent des représentations distribuées pour décrire à la fois le contexte d'une règle et sa production. À l'opposé, les approches décrites par Oakey et Cawthorn (1981) ou Van Coile (1991) visent à extraire des règles qui soient aussi proches que possible de celles qu'un humain produirait. Dans cette optique, la production automatique de règles est considérée comme un moyen d'accélérer la production manuelle de règles, plutôt que comme un substitut. Ceci conduit à se donner une

forme moins rigide pour les règles extraites, incluant par exemple des disjonctions, ce qui contribue à diminuer sensiblement le nombre de règles extraites, et à améliorer leur interprétabilité.

L'architecture de Dietterich et Bakiri

Les travaux de Dietterich et Bakiri se fondent sur l'utilisation exclusive d'ID3. Cet algorithme, capable de construire de manière supervisée des règles de décision binaires, fonctionne dans sa version de base¹⁰ suivant la procédure dichotomique suivante. On dispose initialement d'un ensemble d'objets, catégorisés par une variable binaire, et décrits par un ensemble d'attributs. On commence par déterminer, sur la base d'un critère d'information mutuelle, l'attribut (et les valeurs afférentes) qui permet de séparer au mieux l'ensemble de départ en deux sous-groupes. Ceci détermine un premier niveau de l'arborescence. Puis on itère la procédure pour chacun des sous-groupes, jusqu'à ce qu'ou bien tous les attributs aient été utilisés, ou bien tous les sous-groupes correspondant aux feuilles de l'arbre construit soient homogènes au regard de la variable de catégorisation. Soit maintenant un objet dont on connaît les attributs, mais pour lequel la variable de catégorisation est inconnue : il suffira pour le classer de parcourir l'arbre en testant à chaque bifurcation les attributs du nouvel objet nécessaires pour déterminer la branche suivante à explorer; on répète cette procédure jusqu'à arriver à une feuille de l'arbre de décision, feuille dont la composition dictera le choix de la classe à assigner.

Exposée dans ces termes, l'architecture de Dietterich et Bakiri interprète le problème de la transcription, comme le problème de construire un arbre permettant de classer des graphèmes (les objets), en tenant compte de leur valeur et de celle des graphèmes avoisinants (les attributs qui décrivent les objets). Dans la mesure où le but premier de ces auteurs est de comparer leurs résultats avec ceux de Nttalk, le système de codage des entrées est le même que celui décrit à la section 2.2.1.

Un des mérites de cette approche est de mettre clairement en lumière l'influence de la représentation particulière des phonèmes. Pour utiliser une telle représentation dans leur architecture, ces auteurs sont amenés à construire un arbre de décision *pour chacun des bits* du vecteur représentant les phonèmes, puis à projeter ce vecteur de sortie sur le phonème le plus proche au sens de la distance de Hamming sur les vecteurs de bits. L'architecture complète du système est reproduite sur la figure 2.3.

Quelques raffinements de l'architecture

Décodage par bloc Dans une première expérience, ces auteurs montrent comment améliorer les résultats fournis par cette architecture en intégrant des indices sur la structure morphologique des mots à transcrire. L'idée est relativement simple à mettre en œuvre : durant la phase d'apprentissage, un dictionnaire des terminaisons les plus fréquentes est constitué. À chaque séquence de taille

10. De nombreux raffinements ont été proposés, que nous n'étudierons pas ici.

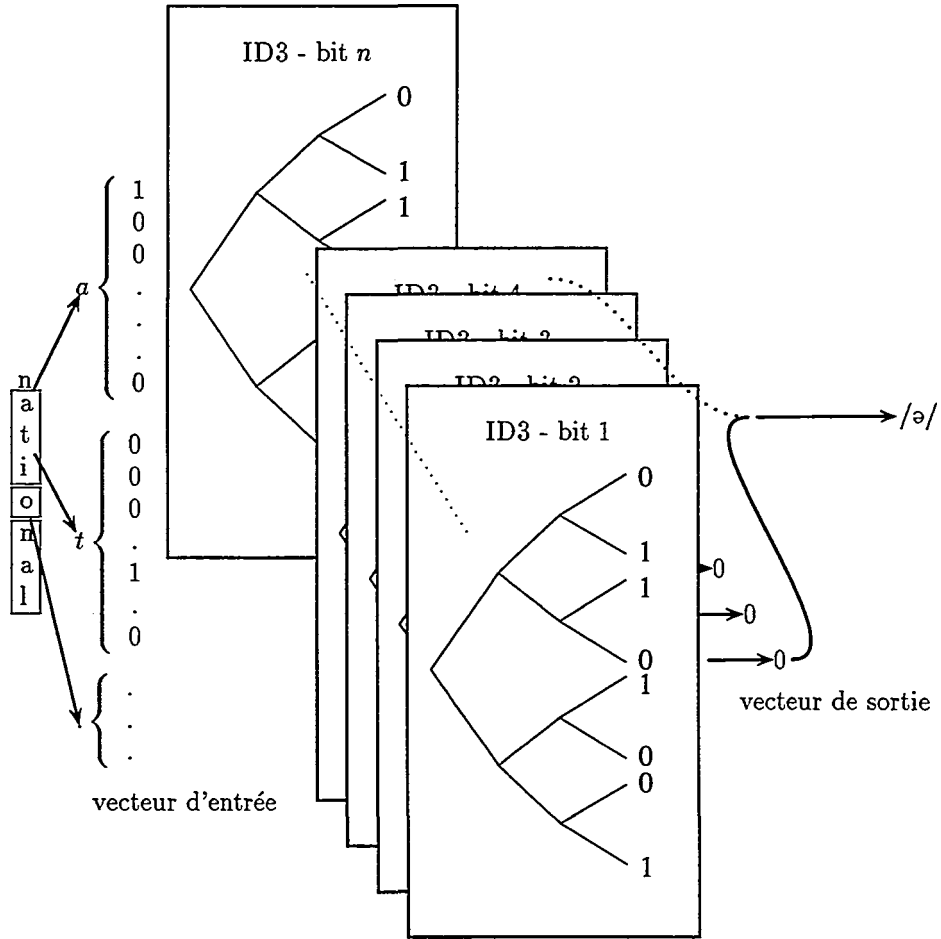


FIG. 2.3 - Architecture de Dietterich et Bakiri. Le graphème à classer est le *o* de *national*, le phonème correspondant un /ə/.

n dans ce dictionnaire est associé un vecteur de bits construit par concaténation des représentations des phonèmes correspondants; ce vecteur binaire a donc $26 * n$ composantes. Ce dictionnaire est mis à contribution durant la phase de décodage du vecteur de sortie par le mécanisme suivant. Pour chaque séquence de taille n dans le dictionnaire, on compare *en bloc* les n vecteurs de sortie correspondant à la finale du mot au vecteur de bits associé à la séquence. Soit ainsi le mot *action* dont on suppose la prononciation inconnue. À chaque lettre est associé un vecteur de bits par application d'ID3. Au lieu qu'ensuite les vecteurs binaires de sortie soient projetés individuellement sur le phonème le plus proche, une comparaison globale portant sur les quatre derniers phonèmes est faite avec le vecteur correspondant à la séquence *tion*, particulièrement fréquente en anglais.

Utilisation de codes correcteurs d'erreurs Un second raffinement particulièrement subtil est introduit par la remarque que la procédure de décodage

sera d'autant plus rarement prise en défaut que les phonèmes seront représentés par des vecteurs de bits bien séparés. Dans le codage de Sejnowski et Rosenberg, chaque bit était interprétable en termes de trait distinctif; ainsi la paire (/p/,/b/) était représentée par des vecteurs distance de Hamming unité, reflétant le fait que ces phonèmes ne se distinguent que par le trait de voisement. Ce que Dietterich et Bakiri montrent, c'est qu'il est préférable d'abandonner cette représentation « naturelle » (pour les linguistes) pour se tourner vers des codes correcteurs d'erreur, i.e. des codes moins sensibles à « petites erreurs ». L'idée est de choisir le codage des phonèmes de telle manière que dans l'espace des vecteurs de bits, deux phonèmes soient au moins à distance de Hamming d . Ceci garantit que si un phonème est prédit avec moins de $d/2$ bits incorrects, la procédure de décodage sera malgré tout capable de le reconnaître sans ambiguïté, en d'autres termes de « corriger » l'erreur. Les méthodes de construction de codes respectant ces contraintes sur les distances sont connues; il est clair toutefois que la taille du code pouvant corriger d erreurs croît avec d (en fait, on peut exhiber de tels codes de taille $2 * d$). Les résultats reproduits dans (Dietterich & Bakiri 1991) sont édifiants sur le succès de cette approche: les performances en termes de phonèmes corrects semblent croître indéfiniment¹¹ avec la taille du code utilisé (la plus grande taille utilisée dans leurs expériences était un code de taille 127). Cette amélioration-là a néanmoins un prix: d'une part le nombre d'arbres de décision à inférer augmente avec la taille du code, et d'autre part l'apprentissage de codes non-fondés linguistiquement conduit à des arbres moins compacts, témoignant d'une difficulté d'apprentissage accrue.

Résultats et discussion

Les expériences de Dietterich et Bakiri visent principalement à comparer leur système avec l'architecture connexionniste de Nettetalk (voir aussi (Mooney *et al.* 1989) pour des résultats convergents); en particulier les lexiques d'apprentissage utilisés dans la plus grande partie de leur tests sont relativement petits (de l'ordre du millier d'exemples). Leurs résultats mettent principalement en évidence le fait que, si Nettetalk obtient des performances significativement meilleures que leur système dans sa version de base, la prise en compte d'informations fréquentielles additionnelles permet de faire converger les performances de ces deux algorithmes. Ainsi, le simple fait d'identifier le vecteur de sortie au plus proche des phonèmes *effectivement rencontrés durant l'apprentissage*, en ne tenant pas compte de ceux qui n'ont jamais été observés, améliore les résultats d'ID3 sensiblement plus que ceux de Nettetalk, et accroît la corrélation entre leurs performances. De même, l'utilisation du décodage par bloc augmente inégalement les performances des deux algorithmes, avec un avantage pour ID3, dont les performances deviennent alors très voisines de celles du réseau connexionniste.

Des expériences complémentaires (Bakiri & Dietterich 1993) visant à confronter cette approche avec un système de transcription par règles (DECTalk) dans des conditions expérimentales plus comparables à celle que nous décrirons dans les chapitres suivants (i.e. utilisant des volumes d'exemples sensiblement plus conséquents) fournissent des points de repères plus fiables.

11. Une limite supérieure existe en théorie, due à la non-indépendance entre les erreurs, mais rien ne permet de la calculer *a priori*.

Ainsi l'utilisation de larges ($n = 127$) codes correcteurs d'erreurs, jointe à un élargissement du contexte graphémique permettant de prendre en compte sept voisins de part et d'autre de la lettre à classer, permet d'obtenir des scores de 64,8 % de mots corrects, et de 93,7 % de phonèmes corrects. L'examen plus fin des erreurs, ainsi que l'évaluation par jugement humain des formes synthétisées conduit même ces auteurs à réévaluer ces résultats, qui approchent alors les 80 % de mots corrects, soit des performances plus conformes aux standards de qualité requis pour une exploitation dans des systèmes de synthèse vocale.

Les travaux de Dietterich et Bakiri sont intéressants à plusieurs titres. Ils permettent d'une part de remettre partiellement en cause les remarques faites à la section 2.2.1 à propos de Nettetalk : la connaissance *a priori* de l'ensemble des phonèmes possible n'est pas nécessairement un avantage, puisque restreindre l'opération de décodage aux seuls phonèmes *observés* améliore en fait les performances. De même, ces travaux jettent un éclairage nouveau sur l'influence de la représentation distribuée des phonèmes : si l'interprétabilité des résultats n'est pas un objectif, alors il semble de loin préférable d'utiliser des systèmes de codage des phonèmes sans pertinence linguistique, mais dotés de propriétés topologiques particulières (par exemple des codes correcteurs d'erreurs).

Ces travaux permettent également de se faire une idée des limites intrinsèques des systèmes de transcription fondés sur des classifieurs. Compte-tenu de la complexité et de la fiabilité relative des données linguistiques manipulées, ainsi que du haut degré de sophistication (décodage par bloc, utilisation de codes correcteurs d'erreurs) des classifieurs employés, il paraît difficile d'obtenir des performances très supérieures à celles mentionnées ci-dessus. En effet, les 93,7 % de phonèmes corrects correspondent à un taux de bits correctement classés voisin de 96 %. Nous aurons l'occasion de développer cet argument dans un contexte plus général à la section 2.4.

2.2.3 Transcription probabiliste

Dans cette section, nous évoquons quelques tentatives qui ont été réalisées pour aborder le problème de la transcription avec les outils traditionnels de la reconnaissance de la parole, à savoir les modèles de Markov cachés (HMM). Une présentation générale et complète de cette classe de modèles excède de loin et nos compétences, et notre propos. Nous nous contenterons de rappeler ici les principes généraux qui fondent cette approche et de discuter quelques uns des résultats qui nous sont connus, en les situant par rapport à la problématique générale de ce chapitre.

Quelques rappels théoriques

Exprimée en termes probabilistes, le problème de la transcription graphème-phonème se ramène à trouver la séquence de phonèmes qui s'apparie le plus vraisemblablement avec une séquence de graphèmes donnée. Il faut pour cela être capable de donner un sens, pour x dans $\mathcal{G}$ à la probabilité (ou la vraisemblance) conditionnelle d'une séquence y de $\mathcal{P}$ sachant x ; puis de savoir résoudre

le programme :

$$y^* = \operatorname{argmax}_{y \in \mathcal{P}} Pr(y | x) \quad (2.1)$$

En utilisant la formule de Bayes, on peut réécrire $Pr(y | x)$ comme dans (2.2) :

$$Pr(y | x) = Pr(x, y) / Pr(x) = Pr(x | y) \times Pr(y) / Pr(x), \quad (2.2)$$

ce qui permet, en utilisant le fait que $Pr(x)$ est constant pour x donné, de faire apparaître deux termes. Le premier, $P(x | y)$ est classiquement interprété comme mesurant la qualité de l'appariement entre les deux séquences de symboles; quant au second, $Pr(y)$, il mesure la qualité de y en tant que chaîne potentielle de $\phi(\mathcal{L})$, et permet de contrôler l'adéquation de y à un modèle phonotactique de la langue. Une transcription est alors d'autant plus probable qu'elle s'apparie plus correctement avec la séquence de graphèmes tout en respectant au mieux l'allure générale d'un mot (phonétique) de la langue.

La modélisation de ce problème, telle qu'elle est réalisée dans (Parfitt & Sharman 1991) ou (Besling 1995), fait appel aux modèles les plus généraux possibles pour décrire des successions de symboles, à savoir à des modèles markoviens. Cette approche, largement validée par ailleurs dans le domaine du traitement de la parole (Rabiner & Juang 1993) comme dans celui du traitement du langage naturel (Charniak 1993), présente de nombreux avantages que nous allons maintenant rappeler.

La modélisation par des modèles de Markov cachés analyse la chaîne de graphème comme le résultat observable du fonctionnement d'un transducteur stochastique. Chaque graphème est émis lors d'un changement d'état dans le transducteur. Dans le cas le plus simple où l'on identifie chaque état à un symbole phonétique, l'émission du mot graphique est tout simplement le fruit d'une traversée du transducteur le long d'une suite d'états (i.e. d'une chaîne de phonèmes). À la différence des modèles markoviens standards, les modèles de Markov cachés abandonnent l'hypothèse que le parcours dans le graphe de transduction est univoquement reconstituable à partir de la chaîne émise : d'un même état, plusieurs transitions peuvent correspondre à l'émission d'un même symbole.

Dans la pratique, la spécification de la connectique du graphe de transduction résulte d'un choix encore plus radical. On fait en effet l'hypothèse qu'il existe une transition de chaque état vers tous les autres, et que cette transition est susceptible de correspondre à l'émission de n'importe quel symbole. Il existe donc systématiquement plusieurs chemins qui peuvent produire un même observable. Une chaîne de graphèmes étant émise, la succession des états qui l'a produite n'est pas uniquement déterminée : elle nous reste *cachée*.

Pour finir de spécifier complètement le modèle, il reste à probabiliser transitions et émissions. L'hypothèse markovienne de dépendance à horizon fini va permettre d'écrire ces probabilités à l'aide d'un petit nombre de paramètres. Pour cela, on suppose que le comportement des processus stochastiques de transition et d'émission en tout état du graphe ne dépendent¹² du comportement

12. Ne nous méprenons pas ici sur le concept de dépendance : postuler une dépendance de

passé du système qu'à un horizon réduit. Ceci permet de décomposer les différents termes de l'équation 2.2 sous la forme de produits d'un nombre réduit de termes, en utilisant la transformation classique suivante (on note n la taille de la dépendance temporelle, et on suppose que les termes correspondant à des indices négatifs disparaissent) :

$$Pr(y = p_1, \dots, p_m) = \prod_{i=1}^{i=m} Pr(p_i | p_1, \dots, p_{i-1}) \quad (2.3)$$

$$= \prod_{i=1}^{i=m} Pr(p_i | p_{i-n}, \dots, p_{i-1}) \quad (2.4)$$

Une décomposition similaire s'applique pour le terme $Pr(x | y)$ qui permet d'exprimer $Pr(x | y)$ comme un produit de probabilités conditionnelles où peuvent intervenir, en plus de l'état courant, des états et/ou émissions passées (dans ce cas, on suppose que la dépendance temporelle vis à vis des émissions passées est de taille n').

$$Pr(x = g_1 \dots g_n | y) = \prod_{i=1}^{i=m} Pr(g_i | p_1, \dots, p_i, g_1, \dots, g_{i-1}) \quad (2.5)$$

$$= \prod_{i=1}^{i=m} Pr(g_i | F(p_i, \dots, p_{i-n}, g_{i-1}, \dots, g_{i-n'})) \quad (2.6)$$

La fonction F de l'équation 2.6 sélectionne un nombre quelconque entre 0 et n des transitions antérieures; entre 0 et n' des émissions antérieures.

Lorsque les paramètres du modèle sont connus ou estimés, il est facile de calculer la probabilité d'émission d'une chaîne de graphèmes quelconque, en effectuant la somme pondérée par la probabilité des séquences d'états qui permettent cette émission du produit des probabilités individuelles d'émission le long du parcours :

$$Pr(x) = \sum_y Pr(x, y) \quad (2.7)$$

$$= \sum_y Pr(x | y) Pr(y) \quad (2.8)$$

Inversement, une chaîne graphémique étant observée, la séquence d'état qui contribue le plus à cette somme est celle qui a le a plus vraisemblablement été parcourue lors de l'émission. Retrouver cette séquence à partir de l'observable correspond à réaliser le programme énoncé par l'équation 2.1, à savoir trouver la séquence de phonèmes (d'états) la plus probable pour une chaîne de graphèmes (un observable) donné. Signalons enfin que pour les deux tâches d'estimation et de décodage, il existe des algorithmes efficaces qui permettent de limiter

taille n ne signifie pas que l'émission d'un symboles ne dépend (au sens statistique) que des n symboles précédents, mais plutôt que la connaissance des n derniers symboles émis contient toute l'information nécessaire à la prédiction du $(n+1)$ ème; et qu'en conséquence l'utilisation de dépendances impliquant un horizon plus lointain sont inutiles, parce que déjà prises en compte dans le modèle.

l'exploration explicite de tous les parcours possibles, les algorithmes de Viterbi (Viterbi 1967) et de Baum-Welsh (Baum 1972).

La réelle difficulté dans ce contexte est celle du calcul ou de l'estimation des probabilités associées aux émissions et aux transitions. Plusieurs cas de figure sont envisageables :

- Dans le cas le plus favorable, on dispose d'une grammaire complète intégrant toute la connaissance nécessaire à la description des chaînes de $\phi(\mathcal{G})$ d'une part, des appariements entre les chaînes de $\mathcal{G}$ et $\phi(\mathcal{G})$ d'autre part. En utilisant ces grammaires en génération, on peut calculer directement et de manière aussi exacte que possible les paramètres du modèle.

Ce cas de figure n'est pas le plus courant : si l'on dispose d'une grammaire complète, on ne cherche pas en général à l'apprendre de manière automatique ! On notera que l'utilisation de grammaires approximatives permet toutefois, et à bon compte, de calculer (ou de simplifier le calcul) d'un certain nombre de paramètres. Pour donner un exemple très simple, si l'on dispose d'une grammaire phonotactique qui restreint la structure des syllabes possibles à la seule forme *CV*, alors on peut en déduire que toutes les séquences de deux voyelles ou de deux consonnes ont une probabilité rigoureusement nulle.

- Dans le cas un peu moins favorable, on dispose de données de supervision, qui fournissent des exemples de paires (suite d'états, émissions associées). En effectuant de simples décomptes sur ce lexiques, on peut calculer directement les paramètres, en estimant tout simplement les probabilités comme des rapports de fréquences. Ce cas de figure est celui dans lequel on dispose de lexiques *alignés*, qui donnent directement les correspondances entre états (phonèmes) et émissions (graphèmes).
- Dans le cas où l'on ne dispose pas de l'information de supervision, il existe des algorithmes itératifs (de type EM) qui permettent malgré tout de réaliser l'estimation.

Dans ce cas, comme dans le précédent, la contrainte de travailler sur des lexiques de taille finie entraîne un réel problème d'estimation, puisque des séquences d'états (ou des émissions) licites peuvent ne jamais être observées, et donc se voir attribuer une probabilité rigoureusement nulle, correspondant à leur fréquence d'occurrence. Il faut dans ce cas faire appel à des techniques d'interpolation (Jelinek & Mercer 1980), en utilisant par exemple pour estimer les probabilités des séquences non-observées des combinaisons linéaires des probabilités de sous-séquences attestées.

Précisons pour conclure que d'autres approches sont, toujours dans un contexte probabiliste, envisageables, qui feraient appel à des représentations plus élaborées des unités modélisées; et notamment celles qui font appel à des grammaires (prosodiques et/ou morphématiques) de mots, dans l'esprit des travaux reportés dans (Hunnicut et al. 1993; Meng, Seneff, & Zue 1994a; 1994c; 1994b). En effet, rien ne s'oppose en principe à ce que l'on applique à la transcription les diverses approches utilisées pour l'apprentissage de modèles

syntactiques (voir (Rajman 1995) pour une revue). À modèles complexes, répond toutefois le besoin de disposer pour l'apprentissage de données complexes (arbres syntaxiques des mots), dont l'élaboration pose de sérieux problèmes, y compris celui de la faisabilité puisque de nombreux faits linguistiques semblent contredire l'existence même de grammaires morphotactiques (voir sur ce point la section 4.2.1). On peut trouver là une des raisons pour lesquelles c'est la modélisation la plus simple qui est la plus fréquemment mise en œuvre. Une autre raison est qu'il est sans doute plus difficile d'exhiber, dans le cas de la transcription graphème-phonème, que dans celui de l'analyse syntaxique, des données linguistiques qui pourraient sérieusement invalider l'hypothèse markovienne de dépendance à horizon fini.

Résultats et discussion

Quelques résultats Les expériences décrites dans (Parfitt & Sharman 1991) se fondent sur l'utilisation de spécifications très simples des dépendances temporelles, puisque $P(y)$ est estimée dans un modèle bi-gramme (i.e. exprimée comme un produit de probabilités conditionnelles de la forme $Pr(p_i | p_{i-1})$); et que $P(x | y)$ est estimé dans un modèle sans passé: les probabilités d'émission ne dépendent que de l'état courant. Ces expériences font appel à des procédures de réestimation, puisqu'elles utilisent les données (non-alignées) du corpus Lancaster/IBM. Dans ces conditions, les résultats obtenus correspondent à 85 % de phonèmes correctement prédits.

Les expériences décrites dans (Besling 1995) sont beaucoup plus complètes, puisqu'elles portent sur trois langues différentes (le français, l'anglais et l'allemand). Les modèles utilisés sont également plus sophistiqués, puisque le modèle phonotactique est estimé par des 7-grammes, et que le modèle d'appariement inclut des dépendances d'ordre 2 concernant aussi bien les émissions que les transitions. Ultime raffinement, des modèles différents sont estimés suivant que les phonèmes sont situés en position initiale de mot, en position finale, ou en position intérieure. Les chiffres les plus intéressants pour nos comparaisons¹³ sont les résultats obtenus sur le lexique CMU, que nous avons également utilisé dans nos expériences. Pour ce lexique, les performances sur un jeu de test sont de 57,2 % de mots corrects, soit 88,7 % de phonèmes corrects. Des résultats sensiblement meilleurs sont obtenus pour des lexiques allemands et français.

Ces derniers résultats nous inspirent deux réflexions contradictoires: de tous les systèmes comparables recensés, c'est sans nul doute l'approche markovienne qui permet d'arriver aux meilleurs résultats (en prenant le lexique NETTALK comme référence); dans le même temps le fort taux d'erreur persistant nous semble indiquer qu'en dépit des raffinements des méthodes d'estimation, ce type de modélisation ne permet pas d'obtenir des performances suffisantes au regard de celles obtenues par des approches à base de règles d'expert.

13. Stefan Besling m'a toutefois signalé en Novembre 95 que des expériences non-encore publiées sur le lexique NETTALK avaient abouti à des résultats de 92,8 % de phonèmes corrects, et de 72,8 % de mots corrects. Ces résultats ont semble-t-il été obtenus en utilisant une version améliorée de l'architecture que nous venons de décrire.

Modèles probabilistes et classification Il est temps pour nous de justifier l'inclusion des modèles probabilistes dans cette section initialement dévolue aux approches classificatoires de la transcription graphème-phonème. Contrairement aux approches décrites dans les deux sections précédentes, la modélisation probabiliste que nous venons d'évoquer ne repose pas, en principe, sur la classification successive des graphèmes à transcrire. L'appariement « un graphème-un phonème », et le pré-alignement des données d'apprentissage qu'il impose n'est pas inscrit en tant que tel au cœur du fonctionnement de ces modèles; d'autres formalisations ont d'ailleurs été envisagées qui permettent des appariements plus complexes (van Coile 1993).

Une autre différence majeure réside dans le fait que les appariements individuels tels qu'ils peuvent être reconstruits dans les modèles markoviens ne reposent pas uniquement sur une série de classifications individuelles, mais sur la maximisation d'un critère qui intègre simultanément tous les appariements qui sont effectués. En ce sens, ces modèles rendent explicite la dépendance entre les correspondances graphème-phonème qui sont successivement réalisées, là où Nettekoven par exemple n'utilisait qu'une dépendance implicite à travers des fenêtres de contexte.

Si, malgré ces divergences profondes, nous avons décidé de classer les approches probabilistes dans cette section, c'est d'une part parce qu'il nous semble qu'elles sont justiciables du même type de critiques que celles que nous adresserons plus loin aux approches classificatoires; et d'autre part parce qu'elles se distinguent à notre avis encore plus profondément des approches du problème de la transcription que nous étudierons à la section 2.3.

2.3 Transcription à base de lexique

Dans cette section, nous examinons une première alternative aux méthodes présentées à la section 2.2. Le parti-pris est ici de bannir toute forme d'apprentissage: en d'autres termes, de fonder la mise en correspondance d'un graphème et d'un phonème uniquement sur l'examen des exemples de correspondances disponibles. En conséquence, le lexique des mots connus est mis à contribution sous sa forme brute pour guider la décision de classification, plutôt que projeté sur un espace de paramètres par le biais d'une procédure d'apprentissage. Assurément, cette manière de faire présente de certains avantages: gain de temps liés à la suppression de l'étape d'apprentissage; plausibilité cognitive en relation avec les modélisations de la mémoire; grande simplicité de mise en œuvre sous ses formes les plus triviales; souplesse accrue due à la « localité » des « règles » extraites; capacité de restitution exacte des prononciations apprises; et ce avec des capacités théoriques d'apprentissage égales à celles des réseaux multi-couches et d'ID3 (Aha, Kibler, & Albert 1991).

Le classement d'un nouvel objet dans le cadre de ce paradigme se fait en deux temps: la première étape est de retrouver dans la base d'exemples le ou les objets qui lui « ressemblent » le plus; la seconde étant d'extrapoler au cas nouveau les assignations de classe de ces objets voisins. Tout le problème ici est d'une part de trouver une mesure de similarité pertinente, qui permette de

retrouver rapidement et avec précision les cas ressemblants; d'autre part de se doter d'un mécanisme d'extrapolation permettant d'effectuer le passage du cas connu au cas inconnu.

Dans ce cadre très général, de nombreuses expériences ont été menées pour traiter le problème de la conversion graphème-phonème ou d'autres problèmes afférents. Dans la première section, nous présentons le système MBRtalk (Stanfill & Waltz 1986; Stanfill 1987), qui a le mérite de poser sérieusement le problème du calcul des (dis)similarités entre objets, et de fournir des performances sur le lexique NETTALK qui sont directement comparables aux nôtres. La deuxième section est consacrée à l'examen du système ANAPRON (Golding 1991; Golding & Rosenbloom 1991), qui s'inscrit dans le paradigme du raisonnement à base de cas. On peut à juste titre considérer que la présentation de ce système, qui intègre un grand nombre de connaissances d'expert, est quelque peu incongrue dans cette revue des méthodes fondées sur une extraction *automatique* de connaissances. Il nous a paru néanmoins nécessaire d'y faire référence, dans la mesure où ce système est d'une part spécifiquement dédié aux noms propres, et où d'autre part il met en œuvre une conception de l'analogie particulièrement riche.

On trouvera par ailleurs un autre exemple d'application du raisonnement à base de cas à la transcription graphème-phonème dans (Lehnert 1987); des mises en œuvre de techniques inspirées de la méthode des k plus proches voisins pour la prédiction de la position des accents lexicaux dans (Daelemans, Gillis, & Durieux 1994a); ainsi qu'une comparaison entre k plus proches voisins et l'algorithme analogique de Skousen (Skousen 1989) dans (Daelemans, Gillis, & Durieux 1994b).

2.3.1 MBRtalk

Calculs de ressemblance

MBRtalk est un système de transcription graphème-phonème qui repose principalement sur une compétence particulièrement aigüe pour reconnaître dans un problème nouveau un « air de famille » avec des problèmes déjà résolus. Étant donné un lexique de mots alignés avec leur transcription phonétique, MBRtalk analyse ce lexique comme une collection d'objets de la façon suivante. Pour chaque lettre de chaque lexie, un nouvel objet est créé, qui est décrit par la valeur de la lettre courante, des lettres qui l'entourent, ainsi que par le phonème et le marqueur prosodique associés¹⁴.

Chaque lettre d'un mot dont la prononciation est inconnue peut de fait être décrite par le même système de variables, à ceci près que la valeur de deux de ces variables (le phonème et le marquage prosodique) est inconnue. La première étape consiste à trouver dans la base d'exemples le ou les graphèmes les plus proches de l'objet à classer. Une première option est d'utiliser la métrique fondée

14. Stanfill et Waltz rapportent en fait deux expériences, l'une pour laquelle l'environnement graphémique est constitué comme dans Nettalk d'une fenêtre de taille 7; l'autre pour laquelle cet environnement graphémique est étendu pour englober jusqu'à 4 caractères avant la lettre courante et 10 après, pour laquelle les 4 derniers phonèmes prédits sont également pris en compte.

sur le nombre de caractéristiques que deux objets partagent. Dans la mesure où toutes les lettres de contexte graphémique ne sont pas également pertinentes pour classer le graphème courant (sur ce point, voir aussi la section B.3), il est manifeste qu'une telle stratégie va conduire à considérer comme voisins des objets (lettres) qui ne peuvent avoir la même prononciation : ainsi le *f* de *fill* et le *b* de *bill*, au titre que les contextes graphémiques autour de ces deux lettres sont identiques.

Une stratégie plus adéquate est d'intégrer l'importance relative des descripteurs dans le calcul de la similarité, par exemple en pondérant chaque caractéristique par la quantité d'information que sa connaissance apporte à la prédiction du graphème à classer. Deux objets seront alors d'autant plus proches qu'ils partagent plus de caractéristiques *pertinentes*. Cette amélioration permet de mettre en évidence la flexibilité introduite en travaillant directement sur les données d'origine. En effet, le calcul de la pondération relative des caractéristiques est ici toujours relatif à la donnée d'un nouvel exemple à classer. De nombreux exemples montrent que suivant la valeur du graphème à prédire, l'élément de contexte essentiel à sa prédiction est susceptible de se trouver soit en aval, soit en amont, et à une distance variable. Différer le calcul des pondérations à un moment où la lettre à transcrire est déjà connue permet de construire une hiérarchie *locale* des caractéristiques, et de résoudre ainsi à bon compte un problème très voisin de celui du choix du contexte dans DEC (voir à la section B.3.1).

Stanfill et Waltz observent à juste titre qu'une métrique ainsi bâtie est encore trop rigide, dans la mesure où elle intègre de manière binaire les ressemblances entre caractéristiques. Celles-ci sont soit égales, soit différentes. Or, il existe des proximités « naturelles » entre graphèmes, dont la prise en compte est susceptible de mener à la construction de voisinages correspondant mieux à l'intuition. L'exemple qu'ils fournissent, concernant la prononciation de la lettre *g* est à cet égard révélateur : si le graphème à transcrire est un *g* suivi d'un *i*, alors il est préférable de le rapprocher d'objets correspondant à la séquence *ge* plutôt qu'à la séquence *ga*. Dans ce contexte précis, et dans ce contexte précis seulement, il faudrait donc pouvoir prendre en compte qu'un *e* ressemble plus à un *i* qu'à un *a*.

Cette intuition est capturée par le procédé suivant : dans un premier temps on exclut de la base d'exemples ceux dont on a de solides raisons de penser qu'ils ne sont pas pertinents. Cela est relativement simple dans ce cas précis, puisqu'il suffit d'ignorer tous les exemples qui correspondent à la prédiction d'une lettre différente de celle en cours d'examen. En comparant sur la base d'exemples réduite les prédictions associées aux différentes valeurs d'une caractéristique, il est alors possible d'en déduire des coefficients mesurant les proximités entre ces valeurs, coefficients qui vont servir à affiner le calcul des proximités entre les cas.

Le processus de transcription d'un mot inconnu dans MBRtalk peut alors être entièrement décrit comme suit. Pour chaque lettre à transcrire, on estime tout d'abord les paramètres de la métrique locale associée; on détermine ensuite dans la base d'exemples les *n* plus proches voisins; un vote est enfin organisé entre ces voisins pour choisir le phonème à prédire, chaque prédiction étant fonction de la somme des distances des voisins qui la soutiennent.

Performances et discussion

Stanfill et Waltz décrivent plusieurs expériences. Dans une configuration où le contexte pris en compte est sensiblement supérieur à celui pris en compte dans les évaluations de Nettek (voir la note ¹⁴ de la page 68), le pourcentage de phonèmes correctement prédits à partir d'un lexique d'environ 4 500 mots est de 86 %, soit 43 % de mots corrects. Même si ces résultats méritent d'être tempérés par ceux du test d'écoute mis en place par ces auteurs, et qui révèle qu'en fait seulement environ 32 % des mots prononcés par MBRtalk sont véritablement intelligibles, ils n'en restent pas moins relativement décevants.

Il faut à notre avis chercher la raison de ce semi-échec dans l'hypothèse qui sous-tend la démarche de Stanfill et Waltz, et qui revient à supposer que la fonction qui met en correspondance exemples et phonèmes exhibe localement (sur les voisinages qu'ils définissent) des propriétés de régularité. Cette hypothèse est manifestement incompatible avec les données linguistiques manipulées, qui se caractérisent au contraire par une forte proportion de « singularités » : existence d'homographes hétérophones, forte lexicalisation des exceptions, etc.

Notons finalement que l'approche de MBRtalk, si elle permet de mettre en œuvre des solutions « naturelles » à certains problèmes auxquels se heurtent les méthodes fondées sur une étape d'apprentissage, se révèle extraordinairement coûteuse en termes de temps de calcul. Le traitement de chaque exemple nécessite en effet l'estimation de nombreux paramètres et le calcul des distances entre le nouvel objet et tous les objets de la base d'exemples. Il faut à ce titre noter que MBRtalk a été développé pour fonctionner sur une machine parallèle comprenant 2¹⁶ processeurs, une des raisons pour laquelle d'ailleurs nous n'avons pas envisagé de poursuivre plus avant dans cette voie de recherche.

2.3.2 ANAPRON, ou la transcription à base de cas

L'approche de Golding (Golding 1991; Golding & Rosenbloom 1991) s'inscrit dans une approche légèrement différente, celle du raisonnement par cas (Kolodner 1993). Cette approche met en effet en œuvre une architecture mixte, couplant des modules à base de règles dédiés à des tâches spécifiques, avec un module de raisonnement permettant de tirer parti directement d'un lexique phonétisé. Si seul ce dernier aspect concerne véritablement notre étude, nous avons tenu néanmoins à présenter les grandes lignes de la démarche mise en œuvre par Golding, parce qu'elle est à nos yeux la plus aboutie des tentatives pour appliquer des méthodes fondées sur l'utilisation de lexiques au problème de la transcription graphème-phonème de noms propres. À cette occasion, nous insisterons tout particulièrement sur la conception de l'analogie qui est au cœur de l'architecture de Golding, en montrant toute la richesse : d'une certaine manière, et par des moyens beaucoup plus frustres, ce sont de telles analogies que nous aimerions pouvoir extraire directement de nos lexiques.

Transcription graphème-phonème dans ANAPRON

L'architecture (complexe) mise en œuvre par Golding est à la mesure de la difficulté de la tâche de transcription des noms propres. Cette architecture comprend en tout huit modules spécialisés, qui sont impliqués en cascade dans la prononciation des mots inconnus. Ces modules sont les suivants :

- normalisation orthographique;
- identification du ou des langage(s) dont le nom est potentiellement issu;
- analyse morphologique;
- transcription graphème-phonème;
- identification de la structure syllabique;
- identification des voyelles accentuées, et des niveaux d'accentuation correspondants;
- sélection de la meilleure solution;
- normalisation phonologique;

Les modules d'entrées et de sortie sont un peu à part, dans la mesure où ils fonctionnent uniquement sur la base de règles, qui sont pour l'un des règles de normalisation orthographique (décomposer les noms composés, uniformiser l'utilisation des majuscules et des minuscules, etc.); et pour l'autre des règles phonologiques obligatoires permettant de passer d'une représentation assez abstraite à une représentation plus proche de la sortie phonétique. À l'inverse, les six autres modules présentent l'intérêt de faire appel simultanément à des connaissances d'expert représentées sous la forme de règles, et à un module capable de tirer partie, en construisant des analogies pertinentes dans un domaine, d'un lexique de 5 000 noms propres dont la prononciation, l'origine, la structure morphologique,... sont connues.

Il est important de noter que ces modules présentent également la caractéristique d'être très prudents dans leurs évaluations, ce qui les conduit à proposer parfois plusieurs interprétations (de faire l'hypothèse de plusieurs langues d'origine, de construire plusieurs analyses morphologiques, etc.) pour un même mot. Ces diverses hypothèses sont propagées en parallèle dans les modules suivants de la chaîne de traitement, dont on notera également qu'ils tirent largement profit des traitements effectués en amont (par le biais notamment de règles de transcription spécifiques aux différentes langues). Ceci permet de différer le choix de la solution à produire jusqu'au moment où une solution complète est disponible. Cette tâche revient à un module de sélection dont le principal intérêt est d'évaluer la plausibilité des transcriptions sur la base de leur forme globale, et de garantir par là même la cohérence de solutions construites essentiellement sur la base de prédicteurs locaux.

Tous ces modules fonctionnent sur le même principe : un jeu de règles d'expert est d'abord mis en œuvre pour proposer une solution, les exemples du lexique étant utilisés pour critiquer ou confirmer la solution courante, voire

pour contribuer à la construction d'une solution alternative. Ce fonctionnement est exemplifié par la description du module de transcription dans les lignes qui suivent.

L'extraction d'analogies dans ANAPRON

Pour illustrer la mise en œuvre du raisonnement analogique dans ANAPRON, nous allons nous appuyer sur l'exemple particulier du module chargé d'effectuer la transcription graphème-phonème. Les mots sont traités graphème après graphème¹⁵, de la gauche vers la droite. Pour chaque graphème à transcrire, le système commence par émettre une prédiction fondée sur l'application d'une règle de réécriture contextuelle. Notons que les règles mises en œuvre lors de cette étape sont très générales, et reflètent simplement les correspondances graphème-phonème les mieux attestées.

Dans un deuxième temps, la procédure extrait du lexique d'exemples tous les mots auxquels la règle sélectionnée s'applique, en prenant soin de distinguer les exemples positifs, qui correspondent à des applications réussies de la règle, des exemples négatifs, pour lesquels la règle échoue à prédire le phonème correct. Reste alors à examiner si le graphème à transcrire est plutôt similaire aux exemples positifs ou aux exemples négatifs. Dans ce dernier cas, la production phonémique associée à la règle est remise en cause; et une règle alternative est construite sur la base des productions attestées. Cette fonction d'extraction et d'évaluation d'analogies est assurée par un module spécifique, dont nous allons maintenant esquisser les grandes lignes du comportement.

Dans l'approche de Golding, la mesure de l'analogie met en jeu simultanément la ressemblance locale et globale entre le mot à transcrire et les mots du lexique. Pour le calcul de ces deux similarités, la démarche est la même: il s'agit tout d'abord d'extraire la règle qui sous-tend l'analogie, puis d'évaluer la généralité et la qualité de cette règle. On notera que pour assurer le calcul de la généralité d'une règle, des classes « naturelles » de graphèmes (voyelles, consonnes, etc) sont pré-définies et représentées dans le moteur d'analogies.

L'exemple suivant (tiré de (Golding 1991, p 78)) décompose le processus d'extraction de la règle locale sous-jacente à l'analogie entre les noms *Fink* et *Blankman*. Le graphème à transcrire est le *n* de *Fink*. Notons enfin que *Blankman*, suite à l'analyse morphologique, est représenté sous la forme *Blank+man*, où le signe *+* isole les deux morphèmes détectés. En unifiant progressivement les contextes autour du *n* dans ces deux noms, le module d'extraction de règles locales va remarquer la présence dans les deux chaînes d'un *k* suivi d'une fin de morphème après le *n*, ainsi que d'une voyelle avant le *n*. La présence d'un groupe de deux consonnes précédant le premier *a* de *Blankman*, alors qu'il n'y en a qu'une avant la voyelle de *Fink*, termine l'expansion du contexte de la règle. La règle sous-jacente à l'analogie locale de ces noms est alors: $n \rightarrow \eta / V _ k\#$. Cette règle se voit affecter deux scores: d'une part une mesure de sa précision qui est fonction de la taille du contexte; d'autre part un indice de sa généralité qui dépend du pourcentage de prédictions correctes de cette règle lorsqu'on l'applique aux autres exemples connus. La mesure d'analogie globale fonctionne de

15. Néanmoins, certaines règles de réécriture transcrivent simultanément plusieurs graphèmes.

manière très similaire, avec toutefois des classes de graphèmes et des conditions d'arrêt plus lâches de façon à pouvoir toujours unifier intégralement dans une règle deux morphèmes.

Suivant la ressemblance ainsi établie entre le nouveau cas et les exemples disponibles, le système choisit alors soit de valider la production fournie par application de la règle initiale, soit d'appliquer une des règles sous-jacentes localement construite.

Performances et discussion

Avant d'entrer dans le vif du sujet, il nous semble important de préciser que les résultats rapportés par Golding ont été obtenus dans le cadre d'un protocole expérimental très différent de celui que nous avons utilisé. Dans ce protocole, toutes les transcriptions, y compris celles résultant de l'interrogation d'experts humains ont été converties au format d'entrée du synthétiseur DECtalk, et évalués à l'aune de tests d'écoute des prononciations synthétisées. Ceci rend les chiffres qui suivent difficilement comparables aux nôtres.

Les résultats annoncés par Golding pour son système sont très bons, et le situent légèrement en dessous des meilleurs systèmes commerciaux (DecVoice ou Orator, dont les modules dédiés aux noms propres sont décrits respectivement dans (Vitale 1991) et (Spiegel & Machi 1990)), puisque le taux de succès face à des noms propres inconnus est de l'ordre de 86 % d'accord avec un groupe d'auditeurs. Ce score est à rapprocher du score de deux experts humains, qui, confrontés au même test, parviennent à des taux de réussite de l'ordre de 93 %; ce qui signifie qu'environ 7 % des prononciations qu'ils ont produites sont jugées parfaitement inintelligibles par le groupe d'évaluateurs.

Cette brève description d'ANAPRON nous a permis de voir comment les idées en germe dans MBRtalk pouvaient être mises en œuvre dans un contexte plus ambitieux, qui intègre dans son architecture tous les composants nécessaires à la production de transcriptions précises des noms propres. ANAPRON nous a également permis de voir, dans la pratique, comment la prise en compte d'informations auxiliaires, via la définition de classes dans l'espace des graphèmes, permet de construire des analogies qui vont au-delà des ressemblances de surface entre les chaînes. Ceci permet entre autre de se fonder sur des lexiques d'exemples plus petits, réduisant ainsi le temps nécessaire à l'extraction de voisins pertinents.

Le second intérêt principal de cette approche mixte réside dans la démonstration qu'il est possible d'utiliser avec succès des formes de raisonnement analogiques dans un contexte où elles sont couplées avec d'autres formes de représentation des connaissances (ici des règles). Cet aspect des systèmes automatiques de transcription, que nous avons pour l'instant laissé de côté, est en fait de la plus haute importance; et il nous semble que la pertinence des approches d'extraction de connaissances pour la transcription, dans la mesure où aucune ne parvient à faire la preuve qu'elle permet d'obtenir des résultats comparables aux meilleurs systèmes à base de règles d'expert, dépend énormément de cette capacité-là. Il est notoirement possible d'introduire des connaissances symboliques dans les systèmes probabilistes, par exemple sous la forme de grammaires; nous avons

évoqué comment à la section 2.2.3. Il semble ainsi qu'il soit également possible et avantageux de réaliser ce couplage entre connaissances d'expert et extraction automatique de connaissance dans le cadre d'approches à base d'analogies.

2.4 À propos des approches classificatoires

À l'issue de ce tour d'horizon, nous développons dans cette section une série d'arguments critiques opposables aux approches classificatoires dans le cadre du problème de la transcription graphème-phonème. Plus que sur les modèles de compétence ou de performance sur lesquelles ces différentes démarches reposent, ou sur le type de modèles informatiques qu'elles mettent en branle, nous articulons principalement notre critique autour du questionnement de l'idée d'une transcription lettre-à-lettre, en investigant successivement les conséquences de cet *a priori* en termes des performances qu'elle permet d'obtenir (sections 2.4.1 et 2.4.2), y compris par passage à la limite (section 2.4.7), des ambiguïtés de transcription qu'elle permet de représenter (section 2.4.3), de la modélisation des phénomènes d'analogie (section 2.4.6), ainsi que de l'opacité qu'elle entretient vis-à-vis de l'existence d'appariements complexes (section 2.4.4) et de phénomènes morphologiques (section 2.4.5).

2.4.1 Performances

La première critique concerne les performances de ces approches. Malgré la multiplicité des angles d'attaque par lesquels ce problème a pu être abordé, les performances sur un lexique comme celui de NETTALK restent, à nos yeux, insuffisantes pour que de telles approches puissent être mises en œuvre dans des applications réelles. Ceci est largement confirmé par l'observation des plus récents synthétiseurs de parole. Malgré en effet la disponibilité accrue de lexiques de transcriptions, et le coût important de développement « manuel » de systèmes à bases de règles, c'est cette option qui a été retenue par tous les industriels au moment de bâtir des systèmes de synthèse multilingues (e.g. (Tzoukerman 1993)).

Cette critique mérite toutefois d'être nuancée, et il existe des langues (l'italien en est un assez bon exemple) pour lesquelles les résultats obtenus par les systèmes d'apprentissage sont tout à fait honorables. Le malheur est que pour ces langues-là précisément, le coût de développement d'une base de règles est également très faible, ce qui fait que, dans tous les cas, cette dernière approche est finalement préférée.

Plusieurs arguments peuvent être invoqués pour expliquer l'insuffisance des résultats, telles que l'inadéquation des représentations utilisées, ou la taille insuffisante des données disponibles. Nous développons ces deux points dans la suite (voir les sections 2.4.7 et 2.4.5). Nous suivons ici une troisième ligne de critique, qui a à voir avec la manière dont les transcriptions sont construites. Dans les systèmes classificatoires, chaque mot est prononcé par simple juxtaposition des prononciations prédites pour chacune des lettres. Si l'on ignore les conditionnements implicites entre productions successives opérés par l'utilisation de fenêtres

de contexte se recouvrant, la probabilité qu'un mot soit correctement transcrit est donc simplement le produit des probabilités de succès des prédictions individuelles. Elle diminue donc mécaniquement avec la taille d'un mot. Ceci présente à nos yeux deux conséquences fâcheuses.

En premier lieu, en supposant les erreurs commises sur les lettres sont indépendantes entre elles - ce qui, d'après les calculs que nous avons pu faire, n'est pas une hypothèse complètement absurde -, on a intuitivement besoin pour obtenir des performances de x % en termes de mots corrects, de garantir une performance par phonèmes de l'ordre de $x^{1/l}$ %, où l est la taille moyenne d'un mot. Pour donner un ordre de grandeur, se fixer une correction par mots de l'ordre de 95 % (avec $l = 6$) impose de parvenir à un niveau de correction par phonème supérieur à 99 %. Compte tenu du haut degré de complexité de la classification, et du degré d'incohérence (ou tout simplement d'erreur) des données disponibles, cet objectif n'est pas forcément réaliste pour toutes les langues.

L'autre conséquence fâcheuse est que l'augmentation de la probabilité d'erreur sur un mot avec sa longueur n'a qu'une faible pertinence linguistique. En effet, il est connu que les mots dont la prononciation est la moins prévisible sont les mots très fréquents et les emprunts. Les mots fréquents sont par ailleurs soumis à des phénomènes d'érosion phonétique; ce sont donc souvent des mots courts. Quant aux emprunts, dont l'origine conduit souvent à une prononciation très déviante par rapport aux autres mots de la langue, il n'y a aucune raison qu'ils soient en moyenne plus longs que les autres. Inversement, les mots très longs résultent souvent de dérivations morphologiques, dont les plus productives sont les plus régulières. On attendrait donc d'un système d'apprentissage convenable de la transcription qu'il rende compte de ces faits dans son comportement, plutôt qu'il exhibe des taux de succès inversement proportionnels à la taille des mots.

2.4.2 Correction et Justesse

On peut en fait étendre un peu ce dernier argument de la façon suivante. Les approches classificatoires sont, dans les faits, *omniphones*, c'est-à-dire qu'elles conduisent à des systèmes capables de prononcer indifféremment tout ce qu'on leur donne. Ceci, nous objectera-t-on, n'est qu'un détail d'implantation: il suffit pour résoudre le problème de créer une classe « imprononçable », et de considérer que tout mot pour lequel une lettre est imprononçable l'est dans sa totalité. À notre connaissance, toutefois, ceci n'a jamais été réalisé; et il nous semble que cet oubli apparent est le signe que des difficultés profondes s'opposent à une telle approche. Réfléchissons un instant sur la forme que pourrait prendre ce type d'extension.

Une première possibilité de mise en œuvre se fonde sur l'association d'un indice numérique à chaque prédiction, indice qui témoigne de sa qualité. Un tel indice existe déjà dans de nombreux systèmes, et sert à départager les prédictions alternatives. Il est alors possible, sur la base de la confiance que l'on accorde à chaque prédiction, de se fixer un seuil en dessous duquel un graphème n'est pas classé (prononcé), et le mot afférent, tu. Ou encore, de calculer un score moyen pour le mot, en dessous duquel on préférera renoncer à le prononcer.

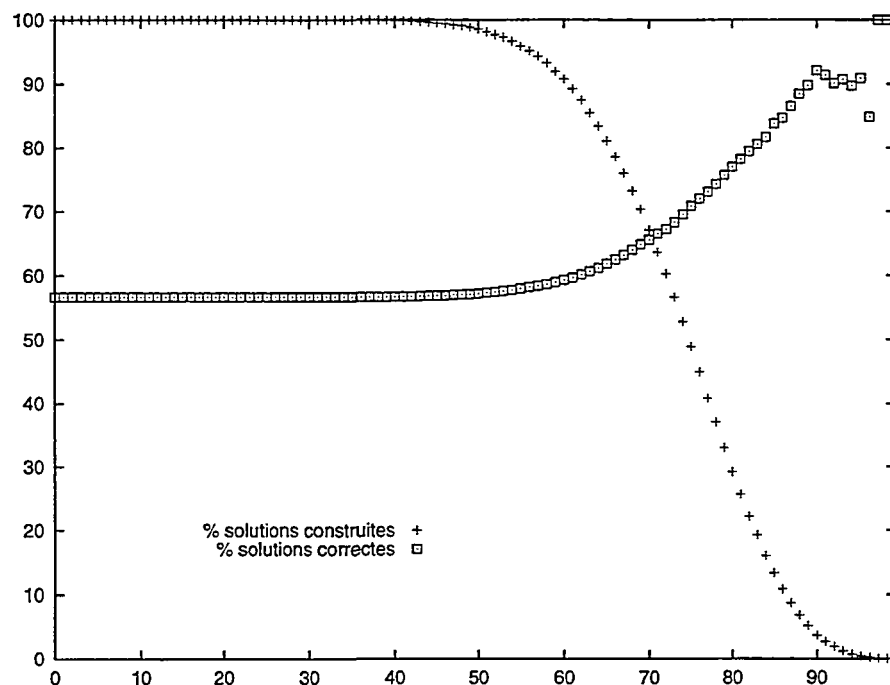


FIG. 2.4 - Taux de réussite de DEC en fonction du score. Les données sont celles de NETTALK

La mise en œuvre d'une telle approche présuppose néanmoins la détermination d'un seuil, tâche qui, dans la pratique est hautement malaisée, et qui trouve peu de justification théorique. Pour donner une idée des difficultés, nous avons associé un tel score à chacune des prononciations fournies par DEC; ce score correspond à la moyenne des scores associés à chacune des règles impliquées dans l'élaboration d'une prononciation. La figure 2.4 représente en fonction du score S , variant de 0 à 100, le pourcentage de mots corrects ayant une note supérieure à S , et le pourcentages de mots ayant une note supérieure à S .

Ce que cette courbe met en évidence, c'est le fait qu'autour de la valeur 70, on observe *à la fois* une augmentation du pourcentage de correction par mot, et une forte diminution de la population de mots dépassant ce seuil de confiance. La question de la détermination d'un seuil de confiance pour ce type de courbe est particulièrement ardue : mettre un seuil supérieur à la valeur pivot 70 conduit à éliminer un trop grand nombre de mots; inversement tout score inférieur à 70 discrimine insuffisamment les mots corrects des mots incorrects.

Cette première approche à base de seuil est en fait largement *ad hoc* : la « règle » de classement d'un graphème à la classe « imprononçable » n'a pas du tout le même statut que les autres. La raison en est qu'elle n'est pas inférée sur le lexique, puisque celui-ci, par nature, ne contient pas d'exemple de cette association.

Alternativement, il est possible d'imaginer de travailler sur des lexiques contenant des exemples de formes, donc de graphèmes, non-prononçables. La dif-

ficulté réelle est dans ce cas de construire de tels mots : le faire de manière efficace requièrerait une grammaire phonotactique de la langue, ce que l'on cherche précisément à extraire automatiquement. Il nous semble pourtant, qu'un système satisfaisant d'apprentissage de la transcription devrait être à même d'extraire, sous une certaine forme, une connaissance de la phonotactique.

Nous voyons plusieurs motivations à la construction de systèmes qui ne soient pas omniphones. Une première raison est que toute chaîne de caractère n'est pas bonne à lire, et que l'humain, en particulier, sait reconnaître les chaînes illisibles et leur appliquer des traitements spécifiques (les épeler par exemple). Ce qu'il est à ce titre souhaitable de mettre en œuvre, ce sont des systèmes qui enregistrent, en même temps que les correspondances graphème-phonème, les agencements spécifiques de graphèmes qui sont permis et donc prononçables, et ceux qui ne le sont pas.

Il existe une seconde raison, plus pragmatique, qui rend de tels systèmes hautement désirables. Si l'on dispose d'un système qui transcrit tout ce qu'on lui présente, la recherche ou le post-traitement des erreurs nécessite l'examen de chaque mot. Si au contraire, le système rejette un certain nombre d'entrées, il est possible d'isoler automatiquement une partie des mots problématiques, et de leur faire subir un traitement particulier. Ceci s'avère spécialement intéressant dans le cadre de la construction semi-automatique de lexiques.

Enfin, mais ce dernier point dépend de la fiabilité du mécanisme de sélection des mots qui ne seront pas prononcés, cette stratégie « laconique » peut permettre à bon compte d'améliorer le pourcentage de mots corrects parmi ceux qui ont été effectivement transcrits.

2.4.3 Les prononciations ambiguës

Comme nous l'avons souligné à la section 1.1.4, un des défis majeurs que doivent relever les systèmes de transcription dédiés à la transcription de mots isolés est celui de la gestion de l'ambiguïté et des transcriptions multiples. Cette ambiguïté a, rappelons-le, de multiples sources (voir également les sections 1.1.2 et 1.1.4) : en plus des variations phonétiques conditionnées, mentionnons le problème des homographes hétérophones, qui suivant les traits morpho-syntaxiques (*le couvent* v.s. *ils couvent*) ou sémantiques (*les fils du disjoncteur* v.s. *les fils du directeur*) qu'ils portent, se prononcent différemment; celui également des formes pour lesquelles plusieurs prononciations sont systématiquement possibles, que cette variation libre reflète une absence de consensus quant à la façon dont ils doivent être lus (emprunts, nom propres, par exemple) ou non. Si l'on peut à juste titre penser que les premiers types d'ambiguïtés, peuvent, par des traitements additionnels (nécessairement *ad-hoc*, pour ce qui concerne la désambiguïsation sémantique), être levés; le traitement des éléments de la seconde famille est du ressort du système de transcription lui-même.

Les systèmes qui fonctionnent par classifications successives des graphèmes sont, sans nul doute, capables de proposer plusieurs classes possibles pour chaque lettre, en associant aux décisions de classement des indices de confiance. Les réelles difficultés commencent lorsque plusieurs graphèmes sont ambigus dans le même mot. En l'absence de mécanisme explicite de conditionnement des pro-

Corpus	% mots	% symboles
NETTALK	2,410	73,910
NETSTRESS	2,800	73,843
BDLEX	1,805	70,068
CMU-20	0,003	64,420
ENST-20	2,975	69,053
ILC-20	11,470	81,198
TUB	0,312	69,663
KUN	4,113	74,092

TAB. 2.1 - *Évaluation d'une stratégie aveugle au contexte.*

ductions successives, l'ensemble des solutions généré par combinaison de toutes les décisions de classement risque fort de contenir des transcriptions plus qu'improbables. Une instanciation précise de ce problème est discutée dans la présentation de DEC (voir page 226 et suivantes).

L'exemple d'un morphème pour lequel plusieurs prononciations sont licites, tel *stein*, qui peut être lu de multiples manières (/stē/, /stēn/, /ʃtajn/, ...), permet de mesurer la difficulté du problème. Pour trois des cinq graphèmes de ce segment (s, i et n), deux classes sont systématiquement possibles (respectivement /s/ ou /ʃ/, /-/ ou /j/, et /n/ ou /-/); et pour la lettre e, l'hésitation est permise entre trois prononciations différentes (/ē/, /e/ ou /a/). La combinatoire complète des classifications individuelles conduit au total à envisager 18 façons différentes de lire ce segment, dont la grande majorité sont plus qu'improbables (/stej/, par exemple).

Considérer la segmentation du mot en graphèmes comme donnée conduit donc ici à une situation inextricable, sauf à introduire, comme dans DEC, des mécanismes plus ou moins naturels, de conditionnement.

2.4.4 Alignement et transcription

Les approches classificatoires reposent crucialement sur la donnée de lexiques dans lesquels les transcriptions phonétiques sont alignées sur les chaînes graphiques associées (c.f. la section 2.1.2). Comme le note toutefois Bullinaria (1995, p. 104) à propos de Nettetalk :

« (...) this degree of pre-processing of the training data leaves very little of the original problem for the network to do (...) »

Il est difficile de savoir jusqu'à quel point cette assertion est vraie. Afin toutefois de nous faire une idée, nous avons évalué un algorithme très simple, dont le mécanisme se fonde uniquement sur la donnée des alignements. Le principe de ce programme est le suivant : il enregistre sur le lexique d'apprentissage l'association lettre-phonème la plus fréquente, association qui est ensuite systématiquement utilisée lors de la transcription d'un mot inconnu. Les résultats auxquels conduit cette stratégie aveugle sont reproduits dans la table 2.1.

On constate à la lecture de ces chiffres que, une fois que l'alignement est construit, la classification correcte d'environ 7 phonèmes sur 10 (avec une légère variation suivant les lexiques) sur des bases fréquentielles est immédiate. La mise en œuvre de techniques d'apprentissage classificatoires plus sophistiquées permet d'améliorer ces scores, et ceci est loin d'être inutile, compte-tenu de la faiblesse des performances en termes de mots corrects de cette approche naïve. Il n'en reste pas moins que la donnée des alignements est, nous semble-t-il, une source d'information précieuse.

2.4.5 Questions de segmentation

Ainsi que nous l'avons exposé à la section 1.2.4, une des compétences essentielle mise en œuvre par l'humain dans l'exécution de tâches de lecture consiste dans la capacité à repérer simultanément dans une chaîne graphique donnée différents niveaux d'organisation étroitement imbriqués, qui structurent cette chaîne en unités plus petites, et qui sont autant d'indices qui facilitent le processus de lecture. La capacité à opérer un découpage en lettres, bien sûr, mais également en phonogrammes, en mores, en syllabes, en pieds métriques, en morphèmes, ou encore à identifier la rime sont autant de facteurs dont l'importance est connue, et qui, à des titres divers selon les langues, font partie intégrante des descriptions linguistiques des processus phonologiques.

Dans ce contexte, la décision radicale de ne considérer qu'un seul niveau de segmentation, celui en lettres, conduit inmanquablement à opacifier l'existence des niveaux de structure supérieurs, qui pourtant sont au principe de régularités particulièrement bien attestées (c.f. par exemple la règle de « Tri-syllabic Laxing » en anglais, ou l'alternance /e-c/ suivant le caractère ouvert ou fermé de la syllabe en français).

Entendons-nous bien sur la nature de l'argument ici développé : le problème tel que nous le comprenons ne porte pas sur les capacités intrinsèques de représentation et de manipulation d'informations symboliques (telles que les marques de début et de fin de syllabe, de début et de fin de morphème, etc.) des mécanismes classificatoires : ces approches sont parfaitement adaptées à la manipulation de ces informations. Nul ne doute par exemple qu'un algorithme tel que DEC saurait tirer tout le parti de chaînes graphiques plus structurées, comportant par exemple les frontières de syllabes et de morphèmes; de la même manière qu'il tire déjà avantage du marquage des débuts et fins de mots. De telles données sont malheureusement rares, et même si elles étaient disponibles, il est difficile de prédire si ces représentations plus complexes amélioreraient ou non les performances des systèmes appris, *à volume d'exemples constant* (voir aussi la section 2.4.7); même si l'exemple de l'application développée par Golding (1991) incite à l'optimisme.

Dans ce contexte, la focalisation exclusive des systèmes classificatoires sur l'aspect appariement de la tâche de prononciation nous semble coupable : les volumes d'exemples disponibles aujourd'hui permettent, nous essaierons de le montrer dans la suite, de s'orienter vers des démarches qui s'attaquent à l'extraction conjointe des segmentations pertinentes et des appariements afférents.

2.4.6 Analogie et plausibilité cognitive

Bien que la plausibilité cognitive des différentes modélisation de l'apprentissage de la lecture soit assez loin d'être notre souci principal, il nous a semblé nécessaire d'introduire ici un semblant de clarification sur le terme d'*analogie*, en relation avec les modèles étudiés à la section 1.2.1. Nous nous contenterons pour l'instant de noter que :

- Les analogies mises en œuvre dans l'approche illustrée par MBRtalk, au même titre que celles extraites par l'algorithme développé par Skousen (Skousen 1989), sont des analogies incomplètes, dans la mesure où elles ne mettent en relation que trois termes, d'une manière que l'on peut gloser par « la lettre *g* est au mot *x* ce que la lettre *g* est au mot *y* ». La même remarque pourrait être adressée aux autres approches classificatoires, en particulier à l'approche connexionniste, dont le caractère analogique, souvent souligné, est rarement formalisé autrement que dans les termes d'une proximité dans un espace métrique, et à ce titre justiciable d'une critique similaire;
- La recherche d'analogies se faisant lettre par lettre, ces approches sont dans la pratique incapables de donner un contenu précis à la notion de *mot analogue* (les mots qui contiennent une lettre analogue?). En particulier, la simulation d'effets de « gang », ou de l'importance de la rime qui sont les manifestations les plus saillantes des processus analogiques dans l'exécution des tâches de lecture (c.f. la section 1.2.3), n'est pas envisageable dans un tel contexte, ce qui vide singulièrement de toute substance la mise en relation entre ces techniques de transcription et les observations expérimentales de Glushko.

2.4.7 La question de la taille des lexiques

Le lien qui unit profondément les approches présentées dans ce chapitre est qu'elles se fondent toutes sur l'utilisation de la distribution fréquentielle des appariements observés dans le lexique d'apprentissage pour élaborer des inférences. Cette caractéristique essentielle est à notre avis à la source de la rapidité avec laquelle des taux de généralisation importants (en termes de phonèmes corrects) sont obtenus. La courbe d'apprentissage de DEC est à notre avis assez exemplaire de cet état de fait : 2 000 exemples suffisent pour obtenir un taux de phonèmes corrects avoisinant 90 %. Des observations similaires sont faites dans (Dietterich, Hild, & Bakiri 1995) pour l'approche connexionniste comme pour l'approche par induction d'arbre de décision : 1 000 exemples permettent d'obtenir des taux de correction par phonème proches de 80 %.

La question qu'il est légitime de se poser de ce contexte est alors celle de l'intérêt d'augmenter la taille des lexiques d'apprentissage (i.e. manuellement annotés), qui sont disponibles dans des volumes toujours croissants (le projet ONOMASTICA fournit d'ailleurs un bon exemple de cette surenchère au niveau de la taille des lexique). Il est en effet acquis que, étant donné un modèle statistique dont les paramètres sont estimés sur un échantillon de taille *N*, la précision de l'estimation (i.e. l'écart-type des estimateurs) augmente dans un rapport de

l'ordre de $1/\sqrt{N}$: c'est tout le sens du théorème central-limite. Compte-tenu de la précision avec laquelle on observe un phénomène, on peut alors déterminer la taille de l'échantillon nécessaire pour obtenir une précision d'estimation du même ordre que l'erreur d'observation; vouloir être plus précis ne sert à rien. Dans le cas de la première classe de modèles étudiés dans ce chapitre (voir section 2.2), le même problème se pose, avec une complication supplémentaire, celle de la mesure de l'influence de la qualité de l'estimation sur les performances en généralisation. On notera que cette interrogation est d'autant plus fondée que les modèles dont il est ici question sont des modèles à mémoire imparfaite, c'est-à-dire qui ne conservent pas explicitement le souvenir de toutes les observations emmagasinées durant l'apprentissage. Pour les autres (voir la section 2.3), s'il est clair qu'un surcroît d'exemples correspond mécaniquement à une amélioration des performances du simple fait de l'apprentissage « par cœur », mentionnons toutefois que, compte-tenu de la complexité des algorithmes qu'elles mettent nécessairement en œuvre, ce gain-là se fait nécessairement au détriment du temps de traitement, qui peut rapidement devenir prohibitif.

En ce qui concerne les systèmes d'apprentissage statistiques, il semble que l'on puisse tenir pour acquis les deux points suivants. D'une part, pour une configuration donnée (nombre de neurones, structure des liaisons dans le cas d'un réseau connexionniste, nombre de descripteurs dans les approches symboliques, ou taille de la dépendance « temporelle » pour les modèles markoviens), il semble qu'il existe une limite au delà de laquelle l'accroissement du nombre d'exemples ne modifie plus que très marginalement les performances en généralisation. De tels résultats sont par exemple observés pour un perceptron multi-couche dans le cas précis de la transcription dans (Ainsworth & Pell 1989) : au delà de 500 000 exemples¹⁶, le réseau étudié ne semble plus tirer partie du savoir additionnel fourni. L'observation de la courbe d'apprentissage de DEC sur deux lexiques de grande taille, reproduite sur la figure 2.5, semble exhiber le même type de comportement : on peut raisonnablement penser que ce handicap est partagé par toutes les approches statistiques.

D'autre part, et ceci a été particulièrement étudié dans les modèles connexionnistes et markoviens, la relation entre qualité d'estimation et capacité de généralisation est complexe. Dans le premier cas, de nombreuses études ont mis en évidence le fait suivant : si on laisse le réseau itérer indéfiniment sur la base d'exemples, il existe un seuil au-delà duquel le gain mesuré en termes d'erreurs commises sur le jeu d'apprentissage est annulé par une dégradation des performances en généralisation (i.e. sur le jeu test) : c'est le problème du sur-apprentissage. En d'autres termes, au delà d'un certain seuil, le système d'apprentissage a « trop bien » appris le lexique d'apprentissage pour pouvoir en tirer des généralisations valables. Le même problème se pose dans des termes légèrement différents dans le cas des modèles markoviens. La relation entre la qualité de l'ajustement du modèle aux données, mesurée par la perplexité du modèle, et sa capacité de généralisation est elle aussi obscure. La difficulté de parvenir par les algorithmes de réestimation traditionnels à de « bonnes estimations » des paramètres reste toutefois le problème majeur, avant celui de savoir si une « bonne estimation » est de toute façon utile ou nécessaire.

16. Chaque exemple correspondant à un phonème, on peut considérer que ce chiffre correspond environ à 55 000/60 000 mots.

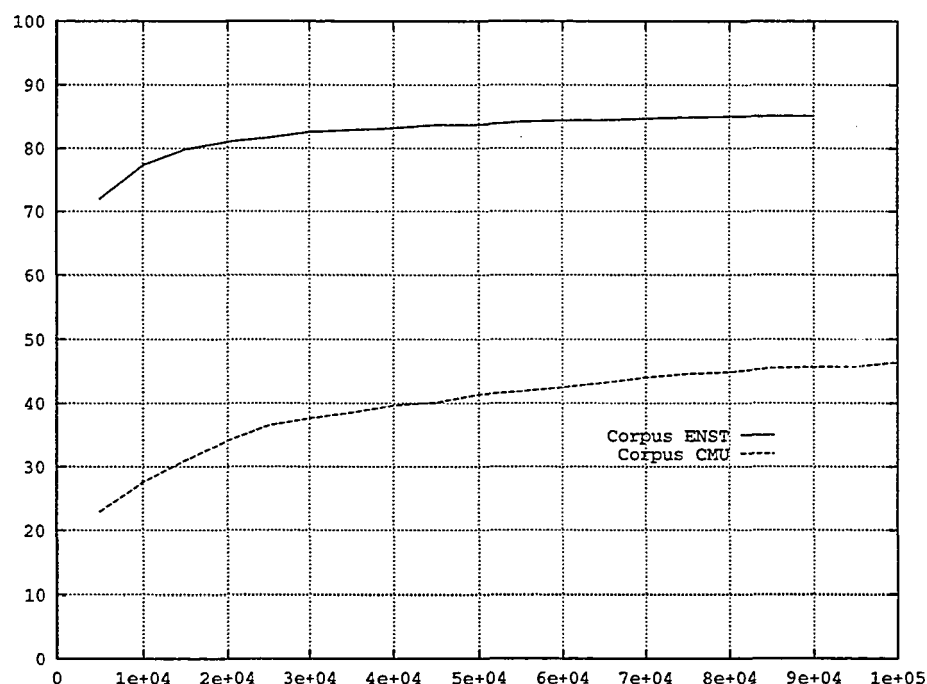


FIG. 2.5 - Évaluation de la vitesse d'apprentissage de DEC sur les lexiques ENST et CMU. Le schéma d'expansion du contexte vaut fb.

Intuitivement, on peut anticiper que les gains à attendre en termes de qualité d'estimation d'un surcroît d'exemples sont, dans le cas précis du langage naturel, bien plus faibles que le taux $1/\sqrt{N}$. La raison est que les données linguistiques déclinent sous des formes variées un type de distribution très particulier, à savoir d'une part des règles très générales massivement attestées; et d'autre part un petit nombre d'exceptions, dont la description requiert toutefois un nombre extrêmement important de règles de faible productivité. Dans ce contexte, on peut légitimement s'attendre à ce que, si l'on choisit avec une probabilité uniforme les exemples supplémentaires, l'amélioration des performances va être simplement proportion de la probabilité d'ajouter des items exceptionnels au lexique d'apprentissage; les formes régulières qui constituent la majorité des nouveaux exemples ne faisant que confirmer des régularités depuis longtemps apprises.

Face à ce problème, deux échappatoires sont possibles. On peut d'une part profiter de l'accroissement du volume de données disponibles pour identifier et apprendre des modèles plus complexes, avec l'argument que les degrés de liberté supplémentaires vont permettre d'apprendre des classes plus grandes d'observables. Dans le cas qui nous concerne, cette complexification des modèles correspond par exemple à la prise en compte de fenêtres de contextes de plus en plus larges. Le doute n'est pas ici permis: augmenter le nombre d'exemples améliore effectivement la qualité de l'estimation des paramètres du modèle, dans le sens trivial où elle permet l'estimation de modèles plus grands. Cette approche est mise en œuvre par Bakiri et Dietterich (1993), qui utilisent des fenêtres de

contexte de taille 15, et, dans une moindre mesure dans (Besling 1995), qui estime des modèles de Markov avec une dépendance d'ordre 7. L'enjeu est ici de pouvoir prendre en compte des phénomènes d'influence à distance de plus en plus grande du graphème source; phénomènes dont on a pu constater qu'ils constituent une des difficultés majeures de la transcription automatique.

Si empiriquement cette approche semble porter ses fruits, et si, en théorie, c'est-à-dire avec des lexiques de taille infinie, elle conduit à une prise en compte correcte de tous les phénomènes observés, il nous semble toutefois nécessaire de s'interroger sur la relation entre le volume de données à ajouter pour pouvoir estimer de manière fiable ces dépendances à distance. Illustrons ce point par l'exemple suivant : supposons que le graphème c_s se transcrit toujours π_s , sauf si il est suivi à une distance d du caractère c_c , auquel cas il se prononce π_c . Dans le cas où d vaut 1, c'est-à-dire où c_c suit directement c_s , il suffira pour un algorithme tel que DEC un seul exemple de cette prononciation pathologique pour apprendre la règle. Si maintenant la distance est 2, il faudra observer toutes les séquences possibles $c_s c_c c_c$, afin de ne pas inférer incorrectement que ce sont certaines configurations données du digraphe cc_c qui influencent la prononciation de c_s : le nombre d'exemples à observer vaut alors N , si N est le nombre de valeurs possibles pour c . On se convaincra aisément en poursuivant ce raisonnement qu'on observe là une progression exponentielle du nombre d'exemples en fonction de la distance. Même si cet exemple ne s'applique qu'imparfaitement aux données linguistiques¹⁷, il permet de mettre en évidence la difficulté pour des stratégies purement statistiques de pouvoir capturer de manière fiable ce type de dépendance à distance.

L'autre possibilité d'échappatoire consiste à opérer une sélection habile des nouveaux exemples (Dagan & Engelson 1995). On pourra par exemple choisir les exemples additionnels dans une procédure incrémentale sur la base de leur capacité à améliorer l'estimation de tel ou tel paramètre important; ou encore parce qu'ils ne sont pas encore correctement couverts par le système en cours d'apprentissage. L'utilité marginale des nouveaux exemples est à ce titre garantie, et il semble que ces techniques fournissent un moyen opératoire pour améliorer la vitesse d'apprentissage.

2.4.8 Conclusion

Nous avons procédé, dans cette partie, à un tour d'horizon assez large des approches qui ont pu être envisagées dans le cadre de l'apprentissage de la transcription graphème-phonème. Nous avons essayé de montrer, en nous fondant sur un algorithme à notre sens prototypique, DEC, que ces approches pèchent en un certain nombre d'aspects qui nous semblent importants : mauvaise gestion des prononciations multiples, comportement à la limite, etc.

S'il fallait résumer d'un mot nos réserves à l'égard de cette classe de mé-

17. On pourra à bon droit m'objecter que dans le cas de la transcription, *toutes les configurations possibles* ne sont pas nécessaires, puisque certaines sont interdites par la phonotactique, et ne s'observeront jamais dans un lexique de test, et que donc ce calcul est fallacieux. À quoi nous répondrons qu'il l'est également à un autre titre, en ce qu'il fait l'hypothèse de dépendances de taille fixe, quand dans la réalité ces phénomènes d'influence à distance, se produisent à des distances *variables*.

thodes, dans le contexte de la transcription, nous dirions qu'elles se fondent exclusivement sur une conception dont nous avons soulignée qu'elle était partiellement erronée du fonctionnement de la correspondance graphème-phonème. En effet, ces méthodes cherchent à apprendre, sur des bases strictement fréquentielles, les déviations par rapport à la situation idéale de notation phonographique, alors que, au moins pour le français et l'anglais, les multiples manifestations (marquage d'oppositions morpho-syntaxiques, préservation de l'étymologie, multiplication des variantes de prononciation pour les noms propres, etc.) de l'autonomie du système graphique incitent à repenser en des termes différents cette correspondance, et à en chercher les régularités en manipulant des unités de taille supérieure au graphème.

Imaginer, dans ce contexte, des techniques à base de données plus respectueuses du fonctionnement réel de la langue, plutôt que forcer les données linguistiques à entrer, tant bien que mal, dans les modèles d'apprentissage élaborés dans d'autres contextes et pour d'autres applications, nous semble indispensable. Ce sont de telles techniques que nous nous efforcerons de présenter dans les deux chapitres suivants.

Chapitre 3

Prononcer par recollement de segments

3.1 Préambule

3.1.1 Introduction

Le chapitre précédent nous a fourni un prétexte pour dresser un réquisitoire contre la conception dominante qui consiste à considérer le problème de l'apprentissage de la transcription graphème-phonème sous l'angle de la classification. Nous allons maintenant présenter et développer quelques unes des voies alternatives à ce type d'approche.

L'idée générale ici mise en œuvre ici consiste à abandonner le principe d'une transcription lettre par lettre, et à envisager la transcription simultanée de larges portions d'un mot inconnu. Le principe général est de *rechercher* des segments¹ du mot inconnu dans le lexique, puis de *recombinaison* les divers morceaux ainsi extraits pour construire une transcription. Pour l'essentiel, notre réflexion s'est focalisée sur les mécanismes de recombinaison, la recherche des segments se faisant de manière très simple, par appariement graphémique direct. Les avantages que nous espérons tirer de cette approche sont multiples :

- le problème de la transcription graphème-phonème est rendu difficile dans certaines langues par la forte dépendance contextuelle de la prononciation des lettres considérées individuellement. Dans les systèmes d'apprentissage classificatoires, chaque graphème x_i d'un mot x inconnu se voit assigner une prononciation $\phi(x_i)$ sur la base de l'observation de cette correspondance dans un ensemble d'exemples. Ce mécanisme interdit dans

1. Une précision de vocabulaire s'impose. Dans la mesure où nous aurons constamment besoin de faire référence dans cette section à des parties de mots, nous utiliserons pour les désigner le terme générique de *segment*. Durant tout ce chapitre, il faudra donc uniquement attacher à ce mot le sens qu'il a traditionnellement en français, en faisant abstraction temporairement de l'autre sens que lui donnent les phonéticiens, à savoir celui de synonyme de *phonème*. Nous utiliserons dans le même sens le terme de *fragment*.

les faits la production d'un appariement qui n'aura pas été observé dans le lexique d'exemples. Rien ne permet d'obtenir la même garantie pour ce qui concerne les segments de taille supérieure. En nous fondant sur l'observation de correspondances entre séquences graphémiques et phonémiques de taille supérieure à l'unité, nous espérons pouvoir limiter les risques de produire des prononciations contenant des séquences par trop improbables;

- ce procédé général conduit à limiter mécaniquement le nombre d'exemples mis à contribution pour la prononciation d'un mot inconnu. Par ce biais, nous espérons également pouvoir donner un contenu plus précis à la notion de « mot analogue »; et éventuellement construire des simulations qui permettent de reproduire certaines des observations faites sur l'humain;
- ce changement d'optique conduit à reposer le problème de l'apprentissage de la transcription graphème-phonème d'une façon qui est à notre avis plus correcte : celui de l'induction d'une correspondance entre deux ensembles de chaînes. Le mot, ainsi que les unités pertinentes linguistiquement (syllabes, morphèmes) ou cognitivement (rime) qui le constituent, sont ainsi replacés au centre des débats. Ce faisant, nous espérons pouvoir faire apparaître des régularités et des ressemblances qui jouent au niveau de ces unités plus larges, dont l'influence dans le domaine de la transcription comme dans celui de la construction d'analogues est largement attestée;
- enfin, et ce dernier point n'est pas le moins important, nous espérons traduire algorithmiquement de manière naturelle le principe *de priorité du structurel sur le fréquentiel*. Nous l'avons en effet noté, l'apprentissage strictement fréquentiel conduit à « l'écrasement » numérique des correspondances exceptionnelles, et par là-même à une incapacité essentielle à traiter correctement les lexies irrégulières. Nous montrerons qu'il est possible d'apporter deux types de réponses à ce problème, en identifiant des critères de décomposition/recomposition *structurels*, qui d'une part donnent un moyen de privilégier des appariements rares, mais structurellement corrects, par rapport à des appariements plus fréquents, mais structurellement moins pertinents; et d'autre part permettent dans une certaine mesure d'identifier automatiquement les éléments trop irréguliers par rapport à un ensemble de formes donné, éléments qu'en dernier recours il nous semble préférable de ne pas transcrire du tout, plutôt que de les transcrire systématiquement de manière erronée.

On notera que l'idée de fonder un mécanisme de transduction (au sens large) d'une séquence sur un principe de recombinaison de (relativement) larges segments de données a également été proposée pour résoudre d'autres problèmes du langage naturel ou du traitement de la parole. Parmi les démarches qui nous semblent les plus voisines de la nôtre, citons principalement l'analyse syntaxique par recollement probabiliste de fragments d'arbres syntaxiques (Bod 1995), ou encore la synthèse de parole par concaténation d'unités de taille variable (Sagisaka & Iwahashi 1995; Lemeur 1996).

Notre travail s'inscrit ici principalement dans la lignée de ceux entrepris par Byrd et Chodorow (1985), et étendus par Dedina et Nusbaum (1991) (voir

également (Pirrelli & Federici 1994c), pour la mise en œuvre d'un mécanisme similaire, nécessitant toutefois une étape préalable d'apprentissage). Cette filiation s'éclaire si l'on considère que ces travaux ont été entrepris en référence explicite aux modèles cognitifs de la lecture par analogie. Il est en conséquence naturel que notre présentation commence par un examen détaillé des capacités de ces travaux, évaluation d'autant plus nécessaire que les mécanismes et algorithmes de transcription originels restent en attente d'une véritable évaluation. En particulier, le modèle de Dedina et Nusbaum, bien qu'unanimentement reconnu comme mettant fructueusement en œuvre des processus d'extraction d'analogies, n'a, à notre connaissance, été évalué que sur une liste de 70 pseudo-mots (dans le cadre d'une mise en application des principes énoncés par Glushko). Nous avons donc été amené à formaliser les intuitions de Byrd et Chodorow dans un algorithme, qui sera présenté et évalué à la section 3.2, et à reprogrammer l'algorithme de Dedina et Nusbaum, afin de pouvoir le soumettre à une batterie de tests plus conséquente (voir l'annexe C). La section 3.3 développe une extension entièrement originale de ces travaux, qui permet une amélioration sensible des résultats précédents.

Une approche alternative, reposant sur le même principe général de recombinaison de segments, est finalement proposée et évaluée dans la section 3.4. La modélisation mise en œuvre dans cette dernière section est de type probabiliste, et diverge sensiblement des objectifs que nous nous sommes fixés, en particulier pour ce qui touche à la transcription (ou l'a-transcription) des séquences graphémiques exceptionnelles.

La comparaison des ces deux réponses possibles à une même préoccupation, celle de la transcription par recombinaison de segments, et leur reformulation dans un cadre formel unifié, nous permettront de dresser, à la section 3.5, un premier bilan critique de ces différentes approches, en reposant notamment le problème de la sélection des unités de segmentation pertinentes pour la phonétisation.

En complément de l'évaluation des diverses techniques présentées dans ce chapitre, notre contribution personnelle concerne plus particulièrement l'élaboration des algorithmes proposés aux sections 3.2 et 3.3, ainsi qu'à la définition du cadre formel unifié présenté à la section 3.5.

3.1.2 Retour sur l'alignement

Les procédures que nous allons décrire fonctionnent à partir de lexiques de transcriptions alignées terme-à-terme avec leur graphie respective. Il est toutefois important de noter que le besoin de disposer de transcriptions alignées est ici largement moins essentiel que dans les approches classificatoires. Ces approches (voir le chapitre 2) requièrent des transcriptions alignées pour respecter la contrainte que toute lettre doit être assignée à une classe; en d'autres termes c'est le parti-pris d'une segmentation lettre-à-lettre qui crée la nécessité de l'alignement. Ici au contraire, c'est pour conserver toute latitude dans le choix de la segmentation que nous avons recours à des transcriptions rigidement alignées. Il est même probable que, nourris de transcriptions renseignées par des alignements plurivoques (voir l'annexe A) plutôt qu'univoques, nos algorithmes se

n°	segment	transcription	n°	segment	transcription
1	chanso	/ʃãsõ/	2	hanson	/ãsõ/
3	chans	/ʃãs/	4	hanso	/ãsõ/
5	anson	/ãsõ/	6	chan	/ʃã/
7	hans	/ãs/	8	anso	/ãsõ/
9	nson	/sõ/			

TAB. 3.1 - Les segments extraits du mot *chanson*. Seuls les segments de taille supérieure à 4 sont répertoriés.

comportent de manière encore plus satisfaisante (voir sur ce point la discussion de la section 2.4). L'exemple du mot *chanson*, développé au début du chapitre précédent, permet d'illustrer ce dernier point. Si l'on se fonde sur l'alignement présenté à la figure (2.1.a) (page 52), on peut en inférer, pour chacun des segments que ce mot contient, une prononciation. Quelques uns de ces segments, ainsi que les prononciations associées, sont listés dans la table 3.1 (les phonèmes « nuls » ont été supprimés).

On mesure bien, au vu du nombre des segments manifestement incompatibles avec la prononciation déduite de l'alignement terme à terme (par exemple les segments 1, 4, 8, et 9), tout l'intérêt que présente alors la disponibilité d'un schème d'alignement plus naturel, dans lequel l'utilisation de symboles « nuls » ne serait plus nécessaire. Considérons à cet effet un alignement alternatif de la lexie *chanson* : $\begin{pmatrix} ch \\ /ʃ/ \end{pmatrix} \begin{pmatrix} an \\ /ã/ \end{pmatrix} \begin{pmatrix} s \\ /s/ \end{pmatrix} \begin{pmatrix} on \\ /õ/ \end{pmatrix}$. La donnée de la correspondance entre *ch* et */ʃ/*, entre *an* et */ã/*, et entre *on* et */õ/* permet en effet de filtrer les segments qui introduisent un point de rupture au sein de ces digraphes, tels que le segment 1 de la table 3.1. Un tel alignement offre donc un moyen de se débarrasser de nombreux segments non-valides, dont la réutilisation ultérieure, indépendamment du mécanisme de recombinaison considéré, a des conséquences parfois fâcheuses. Un premier bénéfice anticipé de ce type d'alignement est donc une amélioration des performances. Un second bénéfice est également anticipé, qui concerne plus spécifiquement les ressources nécessaires au fonctionnement de nos procédures : en filtrant les segments incohérents, on réduit la place requise pour stocker les dictionnaires de segments, tout en améliorant la vitesse de recherche et le nombre de recombinaisons à envisager.

3.2 Transcription par recombinaison binaire de segments

Cette première section est consacrée à la description de la mise en œuvre d'une procédure très simple de recombinaison de *deux* segments en vue d'inférer la prononciation d'un mot inconnu. Cette procédure est schématisée sur la figure 3.1, qui illustre la construction d'une prononciation pour le « mot » *émergence*, en tirant parti des prononciations des mots *énergie* et *émergence*.

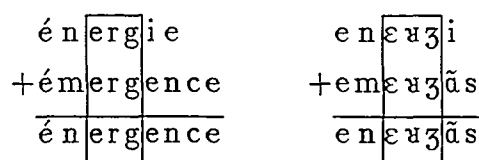


FIG. 3.1 - Un mot plus un mot égale un mot.

Le signe de somme ne doit pas nous tromper : l'opération décrite à la figure 3.1 ressemble en fait plus à un croisement qu'à une addition. Le résultat de l'opération contient en effet un segment qui renvoie à un seul des deux opérandes (le *én* de *énergie*, le *ence* de *émergence*), en plus d'un segment *commun* aux deux opérandes, qui motive et rend possible la recombinaison. On notera que ce procédé est abondamment utilisé dans les créations langagières, qu'elles soient enfantines (« mot-valise »), poétiques ou publicitaires.

Profitions de cet exemple pour introduire quelques définitions, qui seront utilisées de manière récurrente dans ce chapitre. Le mot dont la prononciation doit être inférée sera appelé le *mot cible*; les deux mots que nous allons recombinaisonner pour construire cette prononciation sont les *mots source*. Nous avons par ailleurs encadré sur la figure 3.1 les segments que les mots source partagent, à savoir *erg*, qui appartient aux deux séquences orthographiques, et /εʁʒ/ qui appartient aux deux séquences phonétiques. On utilisera dans la suite le terme de *recouvrement* pour désigner ces segments partagés.

Notre présentation suivra le plan suivant : nous exposons, à la section 3.2.1, l'idée originale de la procédure de prononciation dont nous nous sommes inspiré. Cette version est complétée, formalisée, et enrichie à la section 3.2.2, où nous mettons particulièrement l'accent sur le mécanisme de recombinaison et de classement choisi, et présentons quelques résultats préliminaires. Nous discutons ensuite du problème posé par le silence (section 3.2.3), ainsi de certaines erreurs caractéristiques et des solutions pour les résoudre (section 3.2.4). La section conclusive (section 3.2.5) sera l'occasion de revenir, dans une perspective plus linguistique, sur les mécanismes de prononciation mis en œuvre dans cette section.

3.2.1 L'« algorithme » original de Byrd et Chodorow

Byrd et Chodorow (1985) sont les premiers à proposer un mécanisme inspiré par les modèles cognitifs de lecture par analogie. Tirant les conséquences de la primauté de la rime dans les processus de construction d'analogies, ils proposent le mécanisme très simple suivant pour prononcer des mots inconnus. Étant donné un lexique de prononciation, et un mot inconnu à transcrire phonétiquement, la procédure qu'ils suggèrent comprend les trois phases suivantes :

- *segmentation* : cette étape correspond à l'identification dans le mot cible de deux segments, qui se recouvrent éventuellement;

- *appariement* : le segment initial et le segment final sont recherchés dans le lexique, et les chaînes phonétiques correspondant à ces segments sont extraites. Pour une chaîne graphique donnée, plusieurs chaînes phonétiques peuvent être trouvées dans le lexique;
- *recombinaison* : cette dernière étape consiste à recoller les chaînes phonétiques, de façon à obtenir la prononciation du mot dans son ensemble. Ce processus peut conduire à la construction de plusieurs transcriptions, qu'il est alors nécessaire de classer, sur la base de critères fréquentiels.

La procédure de segmentation effectivement utilisée par Byrd et Chodorow fonctionne pour ainsi dire par défaut : le segment initial (respectivement final) utilisé pour la phase de recombinaison est simplement le plus long segment initial (resp. final) disponible dans le lexique. D'autres procédures de segmentation plus sophistiquées sont proposées par ces auteurs, qui envisagent en particulier de tirer profit de la structure syllabique, mais ces procédures alternatives ne semblent pas avoir été évaluées. Enfin, il est important de souligner que, comme le notent Byrd et Chodorow, il existe des configurations pour lesquelles aucune prononciation ne peut être produite : il suffit pour cela que les plus longs segments disponibles dans le lexique soient insuffisants pour recouvrir complètement le mot cible.

L'exemple suivant (Byrd & Chodorow 1985, p. 281) permettra de mieux comprendre le fonctionnement de ce mécanisme, ainsi que ses zones d'ombre. Dans cet exemple, *brange* est le mot cible. Les plus longues sous-chaînes graphiques initiales et finales trouvées dans le lexique sont respectivement *bran* (21 occurrences) et *ange* (23 occurrences) (l'identification du mot *range*, qui exhibe le plus long appariement final avec le mot cible, ne semble pas stopper la procédure de recherche). Le premier segment est prononcé de façon uniforme dans les 21 mots qui le contiennent : /bræn/; pour le second segment, les prononciations suivantes sont observées : /enj/ (20 occurrences), /ænj/ (1 occurrence), /ɪnj/ (1 occurrence), et /anj/ (1 occurrence). Compte-tenu de ces segments phonétiques, le résultat de l'étape de recombinaison tel qu'il est énoncé par ces auteurs est relativement surprenant puisque seules deux transcriptions alternatives sont proposées : /brænj/ et /brenj/. L'élimination d'une prononciation telle que /branj/ n'est en effet pas réellement justifiée : cette prononciation utilise un segment final (anj) qui pourtant est aussi fréquent que /ænj/, et qui ne partage, comme /enj/, aucun recouvrement phonémique avec le segment initial.

Ce mécanisme, d'une simplicité de bon aloi, semble être un candidat intéressant pour fonder un système de prononciation à base d'analogies. Byrd et Chodorow sont malheureusement muets sur ses performances, aussi bien en termes de mots correctement transcrits qu'en termes de silence. De plus, le fonctionnement exact de la procédure est insuffisamment spécifié, puisque nous l'avons vu, des incertitudes subsistent quant à la nature de la condition d'arrêt de l'étape de recherche d'une part, et quant à la définition précise du procédé de recombinaison et de classement d'autre part, en particulier lorsque des prononciations alternatives sont en concurrence. Nous avons donc été conduit à compléter la formalisation de cette procédure, et à l'implanter sous diverses formes, afin d'en évaluer les réelles potentialités.

3.2.2 Un algorithme de recollement : JUPA

Présentation formelle

La version initiale de cette procédure est décrite formellement dans l'algorithme 1, baptisé JUPA. Dans cette présentation, $\mathcal{L}$ réfère au lexique de mots connus.

Algorithme 1: Prononciation par recollement de deux segments

Entrée: Un lexique $\mathcal{L}$, une chaîne $x = c_1 \dots c_{l(x)}$ à transcrire

Sortie: Un ensemble de transcriptions possibles trié par plausibilité

JUPA(x)

- (1) $\phi(x) \leftarrow \lambda$
- (2) Construire $D(x) = \{y \in \mathcal{L} / (t = pgcp(x, y)) \neq \lambda\}$, trié par valeur décroissante de $L(t)$. La valeur maximale de $L(pgcp(x, y))$ pour $y \in D(x)$ est notée lp_{max} .
- (3) Construire $F(x) = \{z \in \mathcal{L} / (t = pgcs(x, z)) \neq \lambda\}$, trié par valeur décroissante de $L(t)$. La valeur maximale de $L(pgcs(x, z))$ pour $z \in F(x)$ est notée ls_{max} .
- (4) $l \leftarrow lp_{max} + ls_{max} - l(x)$
- (5) **tant que** $l > 0$
- (6) **pour tout** $(y, z) \in D(x) \times F(x) / l(pgcp(x, y)) + l(pgcs(x, z)) = l$
- (7) $lp \leftarrow l(pgcp(x, y))$
- (8) $ls \leftarrow l(pgcs(x, z))$
- (9) $r = c_{lp+ls-l(x)} \dots c_{lp}$
- (10) **si** $\phi_y(r) \neq \phi_z(r)$
- (11) **continuer**
- (12) $\phi(x) \leftarrow \phi_y(c_1 \dots c_l) \phi_z(c_{l+l'-l(x)} \dots c_{l'})$
- (13) enregistrer la solution $\phi(x)$, en incrémentant son score
- (14) **fin**
- (15) **si** des solutions ont été exhibées
- (16) **break**
- (17) $l \leftarrow l - 1$
- (18) **fin**
- (19) **si** des solutions existent
- (20) proposer les solutions par score décroissant

Reprenons l'exemple discuté par Byrd et Chodorow, pour en détailler le traitement dans ce cadre formalisé. Les deux segments les plus longs extraits du lexique sont *bran* (21 occurrences) et *range* (1 occurrence), qui ont trois caractères en commun (*ran*). La valeur de l est donc initialisée (ligne 4) à 3. La prononciation du recouvrement graphique de taille 3 diffère néanmoins dans les parties initiale (/ræn/) et finale (/ren/). La recombinaison sur la base d'un tel recouvrement est donc interdite : c'est le sens du test de la ligne 10). Il faut donc considérer comme base de la recombinaison un recouvrement de taille 2, et donc envisager tous les recollements possibles des segments *bra* et *range* d'une part, et des segments *bran* et *ange* d'autre part. Les données fournies par Byrd et Chodorow ne permettent pas de déterminer toutes les solutions, mais il est

certain qu'au moins la solution /brænɟ/ sera construite à cette étape. Il existe en effet un mot au moins pour lequel la chaîne finale *ange* se prononce /ænɟ/, mot qui peut donc se recombinaison avec ceux qui commencent par /bræn/. La procédure se termine donc pour cette valeur du recouvrement (test ligne 15), puisqu'elle permet la construction d'au moins une solution.

Le rôle de l'alignement

Disposer de données alignées, dans le contexte de cet algorithme, sert uniquement à pouvoir être en mesure de mettre en relation des segments dans les domaines graphiques et phonétiques (i.e. de donner un sens et un contenu à $\phi(r)$, $\phi_y(c_1...c_l)$, etc.).

L'utilisation d'un système d'alignement qui autorise les nuls graphémiques ne pose alors guère de difficultés techniques : il suffit de se doter d'une convention pour l'alignement des segments qui permette d'éliminer ces caractères spéciaux pendant l'étape de stockage du lexique. Reprenons l'exemple du couple (*voy-age*, /vwajaz-/): si le segment initial est utilisé pour une recombinaison, il faut être en mesure de déterminer la chaîne phonémique correspondante, /vwa/ ou /vwaj/. La convention que nous avons adoptée est d'inclure les phonèmes appariés aux nuls graphiques dans les segments gauches pour les chaînes initiales, droits pour les chaînes finales. Selon cette convention, la contrepartie phonémique de *voy* est /vwaj/; et celle de *age* est /az-/.

Une fois une telle convention fixée, la seule conséquence éventuellement entraînée par la prise en compte de graphèmes nuls est une différence de taille des recouvrements dans les domaines graphiques et phonétiques. Ce cas est illustré par la transcription du mot *boxon* par recombinaison de *boxeur* et de *saxon*: au niveau graphique le recouvrement est de taille unité (le *x*), au niveau phonétique, il est de taille 2 (/ks/). Si plusieurs solutions basées sur des recouvrement phonémiques de tailles différentes sont en concurrence, c'est celle qui est fondée sur le plus long qui est préférée.

Un critère structurel de classement

L'exemple de la transcription du mot cible *brange* permet de mettre en évidence la manière dont s'opère la sélection entre les nombreuses solutions qui pourraient être construites à partir d'une telle procédure de recollement. La procédure considère en priorité les solutions fondées sur les plus grands recouvrements, et s'arrête dès qu'une solution est produite, indépendamment de la fréquence des solutions alternatives qui pourraient être construites sur la base d'un recouvrement plus limité. De cette manière, notre algorithme met en application le principe suivant :

la meilleure solution est celle qui est obtenue sur la base du plus grand recouvrement phonétiquement compatible entre les segments initiaux et finaux; les critères fréquentiels ne sont utilisés pour départager les *ex-æquo* pour le critère du recouvrement.

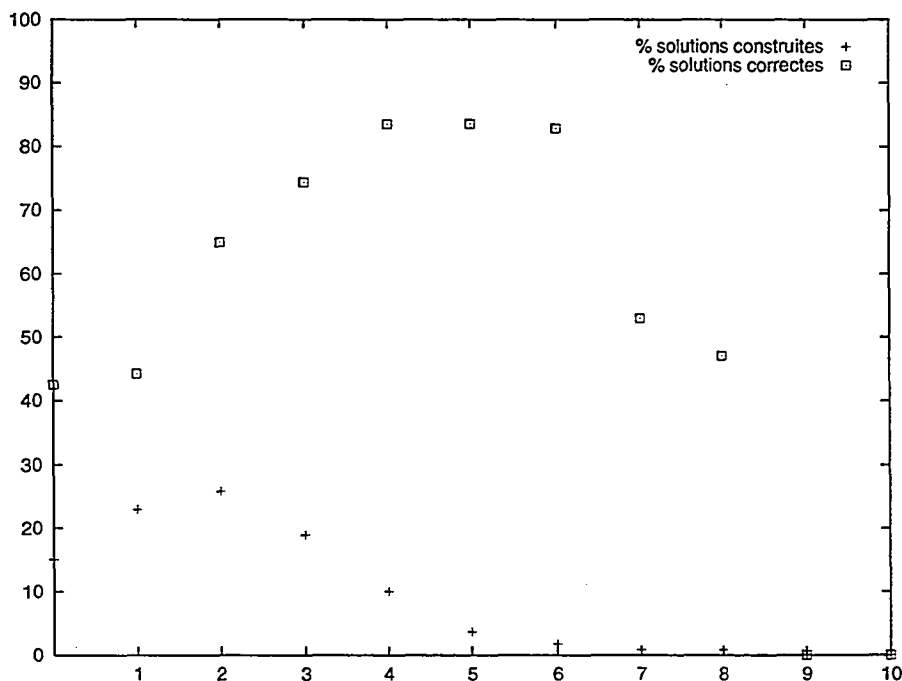


FIG. 3.2 - Taux de correction en fonction de la taille du recouvrement. Les données sont celles du lexique NETTALK.

La primauté accordée à la longueur du recouvrement se justifie de la façon suivante : dans la mesure où la prononciation d'un graphème, et, plus généralement, d'un segment quelconque est hautement dépendante du contexte orthographique dans lequel il apparaît, il est particulièrement important de contrôler les effets contextuels au voisinage de l'endroit où se fait le recollement entre les deux segments sélectionnés. On peut en effet penser que les graphèmes les plus éloignés de cette zone de recollement seront moins influencés dans leur prononciation par la procédure de recombinaison. La manière la plus simple d'effectuer ce contrôle consiste alors à choisir le plus long segment commun aux chaînes initiales et finales tel qu'il existe des mots qui s'accordent sur la prononciation de ce segment. Ce choix se trouve partiellement justifié sur le graphique 3.2, qui représente les pourcentages de mots transcrits de manière exacte en fonction de la longueur du recouvrement.

Ainsi qu'il apparaît sur cette courbe, le pourcentage de mots transcrits avec succès est une fonction croissante de la taille du recouvrement, au moins jusqu'à la valeur 5. Comme le suggère aussi ce graphique, les recouvrements importants, s'ils garantissent des transcriptions de qualité, sont malheureusement disponibles dans des proportions toujours plus faibles. On peut s'interroger sur la décroissance du taux de mots correctement transcrits pour les fortes valeurs du recouvrement. Une raison est manifeste : la rareté de ces configurations ne permet pas d'estimer précisément les chances de produire une transcription correcte. Il existe de fait une autre raison, que l'examen détaillé des erreurs nous

		JUPA		
		Correct	Faux	Total
DEC	Correct	86,59	13,41	100,00
		49,07	7,60	56,67
		79,75	19,76	
	Faux	28,76	71,24	100,00
		12,46	30,87	43,33
		20,25	80,24	
	Total	61,53	38,47	100,00
		100,00	100,00	

TAB. 3.2 - *Matrice de confusion entre les transcriptions de DEC et JUPA.*
Les données sont celles de NETTALK.

révèle, et sur laquelle nous reviendrons plus loin (voir la section 3.2.5).

Résultats préliminaires

Les résultats obtenus en utilisant cette version de base de l'algorithme (voir la table 3.2) sont particulièrement encourageants, surtout si on les compare aux résultats de DEC, notre algorithme de référence.

Le premier enseignement à tirer de la lecture de cette table est que JUPA surpasse DEC de façon sensible en termes de mots correctement transcrits : près de cinq points (pour les taux de mots corrects) séparent les deux algorithmes, cette différence étant significative au seuil de 0,001². Par contre, les algorithmes semblent rencontrer les mêmes difficultés, et buter sur les mêmes mots, puisque près de trois-quart des mots mal prononcés par DEC le sont également par JUPA.

Comme nous l'avons mentionné, la stratégie de recollement est susceptible d'être mise en échec, et de rester muette quant à la prononciation de certains chaînes. Ce taux de silence reste toutefois assez modéré, puisqu'en moyenne, seuls 6 % des mots ne peuvent être complètement recouverts par deux segments disponibles dans le lexique. Pour de telles cibles, JUPA ne construit pas du tout de transcription. La section suivante est consacrée à une discussion approfondie de ces configurations.

3.2.3 Quelques mots sur le silence

La procédure analogique mise en œuvre exhibe ici un comportement inhabituel par rapport aux stratégies classificatoires : certains mots ne sont pas prononcés, ce qui contribue largement à dégrader les performances mesurées en termes de *phonèmes corrects* : quand un mot n'est pas transcrit, tous les phonèmes que sa transcription contient sont faux. Ce qui pourrait passer pour un handicap, une infirmité de notre procédure se révèle pourtant, à l'examen, présenter des aspects positifs.

2. Les procédures statistiques de comparaison d'algorithmes sont décrites dans l'annexe D.

seigneur	smithereens	protuberance
sorghum	cyanide	espadrille
undercrof	broccoli	countinghouse
blossom	cellulosic	connubial
scissors	apocrypha	subservience
cuisine	barracuda	brindled
aileron	bacteriology	nuptial
furbelow	beeswax	bungalow
aardvark	plainclothesman	hustings
qintar	caterwaul	epithalamium
workaday	malevolen	pollywog
kangaroo	beaujolais	pharaoh

TAB. 3.3 - *Quelques mots imprononçables. Les expériences sont conduites avec le lexique NETTALK.*

Il existe en effet des chaînes de caractères qui sont, même pour l'humain, d'une lecture difficile, voire imprononçables. C'est bien sûr le cas des séquences pour lesquelles on ne peut imaginer de transcription respectant les contraintes phonotactiques les plus élémentaires : les séquences sans voyelles en sont l'exemple le plus immédiat. C'est également vrai de séquences qui présentent localement des agencements de graphèmes incongrus, inédits, et pour lesquelles il est particulièrement difficile d'identifier correctement, parmi de trop nombreux candidats, le couplage segmentation/appariement conduisant à la prononciation la plus plausible. Considérons par exemple le nom propre suivant, d'origine polonaise : *Majchrzak*. Ce nom est indiscutablement prononçable, du moins ceux qui le portent sont-ils capables de le faire. Pour un locuteur du français, même familiarisé avec la prononciation de certaines séquences caractéristiques du polonais écrit, un certain nombre d'ambiguïtés subsistent, portant notamment sur l'interprétation à faire des lettres et groupements *j* (/j/ ou /ʒ/), *ch* (/ʃ/ ou /k/), *rz* (/rz/ ou /ʒ/ ou encore /rəz/). En l'absence d'exemples voisins, tout particulièrement d'exemples comprenant le groupe consonantique *jchrz*, cette chaîne est de fait *imprononçable*.

Le mécanisme de recollement mis en œuvre dans JUPA garantit que, s'il ne se fonde que sur un lexique de mots du français, il sera incapable de prononcer les séquences graphémiques entièrement consonantiques. On conçoit aisément en effet que, dès que la séquence à prononcer sera assez longue, les candidats au recollement se feront rares. Il est par contre important de vérifier que ceci est aussi vrai du second type de mots, c'est-à-dire de ceux qui sont atypiques par rapport à un lexique donné. C'est sensiblement ce que l'observation des mots non-prononcés dans nos expériences avec le lexique NETTALK nous permet de constater. Dans la mesure où reproduire la liste de tous les mots non-transcrits était impossible, et où en sélectionner habilement une sous-partie aurait été indelicat, nous présentons *in extenso* dans la table 3.3 la queue de cette liste.

À la lecture de cet inventaire farfelu, il apparaît qu'on peut grossièrement identifier trois grands types de mots imprononçables, à savoir :

- une classe d'emprunts, de *broccoli* à *beaujolais*, en passant par *espadrille*

et *cuisine*, pour ne citer que ceux qui nous sont, à divers titres, les plus familiers;

- une classe de dérivés ou composés morphologiquement complexes, dont *plainclothesman*, *cellulosic* et *workaday* sont de dignes représentants;
- une classe résiduelle, moins fournie, contenant par exemple *blossom* et *nuptial*, dont la présence dans cette liste est plus surprenante.

Il semble donc que JUPA soit en mesure d'identifier, dans une liste de termes à transcrire, ceux qui sont réellement les plus marginaux par rapport au lexique de référence. Le phénomène de silence n'est pas uniformément réparti, mais atteint des familles de mots qui partagent des caractéristiques linguistiques identifiables. On notera que, au regard de la tâche de transcription elle-même, ces classes de mots non-prononcés ne sont pas tout à fait équivalentes : si les emprunts sont notoirement des termes difficiles à transcrire, les dérivés morphologiques complexes eux sont nettement moins irréguliers.

Ce qui passait donc pour une lacune de la procédure se révèle finalement être un aspect plutôt positif de son fonctionnement, puisque JUPA fournit non seulement un mécanisme de transcription, mais également un mécanisme de détection des séquences d'aspect peu familier, séquences dont la prononciation peut être confiée à des modules de traitement différents (par exemple à un système à base de règles, à la manière dont fonctionne TTS (Coker, Church, & Liberman 1990)).

De surcroît, cette sélection ne s'opère ni par le biais de connaissances données *a priori*, telles que, par exemple, des règles d'exclusion phonotactiques, ni par le biais de l'extraction d'un modèle. Au contraire, elle se fonde uniquement sur les connaissances lexicales disponibles à un instant donné, ce qui dote ce mécanisme d'une réelle flexibilité : chaque mot nouveau qui intègre la base d'exemples est susceptible de diminuer le silence, parce qu'il permet d'éclaircir l'algorithme sur la prononciation de mots similaires (i.e. partageant un même segment initial ou final). À ce titre, on est en droit d'espérer que la croissance continue du lexique conduirait à un taux de silence plus faible, corrélatif à un affinement de la classe des mots imprononçables qui, à terme, ne devrait contenir que les emprunts, et les dérivés trop complexes pour être transcrits par un recollement de deux segments. La table 3.4, qui reproduit en parallèle les performances de JUPA en termes de silence, de justesse par mots et par phonèmes pour des tailles croissantes du lexique d'exemples, permet dans une certaine mesure d'observer cette décroissance marginale du taux de silence.

Cette disparition progressive du silence avec la taille du lexique d'exemples est également mise en évidence, sur une échelle plus large, dans d'autres expériences que nous avons conduites. En utilisant par exemple le lexique ENST, qui contient environ 100 000 noms propres, nous avons obtenu la courbe d'apprentissage reproduite à la figure 3.3. Sur cette courbe figurent également les taux de silence et de justesse.

La forme de cette courbe, sensiblement moins plate que la courbe équivalente pour l'algorithme DEC (reproduite en page 233) met en évidence un autre aspect positif de la stratégie de prononciation prudente adoptée dans JUPA. Dans le

Nombre d'exemples	% silence	% mots	% symboles
2000	44.065	47.530	89.279
4000	29.100	52.803	90.853
6000	21.240	56.152	91.743
8000	16.265	58.212	92.257
10000	12.880	60.248	92.715
12000	10.415	61.682	93.008
14000	8.825	63.260	93.318
16000	7.470	64.558	93.593
17940	6.370	65.716	93.822

TAB. 3.4 - Performances de JUPA lorsque le lexique croît. Les données sont celles de NETTALK.

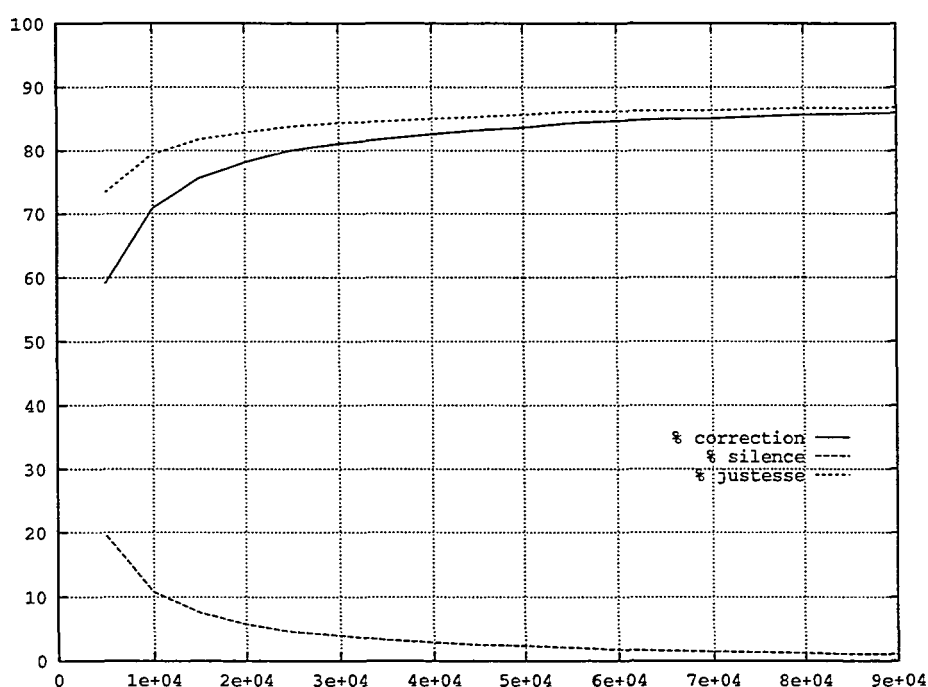


FIG. 3.3 - Courbe d'apprentissage sur le lexique ENST.

contexte d'une approche analogique, les exemples supplémentaires fournis au système semblent ainsi conserver une « productivité marginale » sensiblement supérieure à ce que l'on observe pour les approches classificatoires.

On peut finalement noter que, reconsidéré à la lumière de ces remarques, le mécanisme simple mis en œuvre par JUPA est encore plus efficace qu'il n'y paraît : évalué en termes de *justesse par mot* et de *justesse par phonème*, il conduit, sur le lexique NETTALK, à des scores supérieurs à 65 % et 93 % respectivement, soit près de 10 points (pour les taux de mots justes) de plus que DEC, pour qui justesse et correction se confondent. Ce dernier point est également important : dans une optique applicative, il est clairement préférable de disposer d'un mécanisme qui

prononce peu, mais avec une grande précision, que d'un mécanisme qui transcrit tout, sans qu'il soit possible d'identifier automatiquement les erreurs.

3.2.4 Variations et extensions

Autour de ce modèle de base, un certain nombre d'extensions ont été envisagées, de façon à compenser certaines des faiblesses et lacunes les plus flagrantes de notre procédure. Il existe manifestement deux voies d'amélioration des performances de JUPA : augmenter le nombre de mots transcrits correctement, et diminuer le nombre de mots non-transcrits. Les variations que nous allons développer dans les lignes qui suivent s'attaquent toutes à la résolution du premier problème, et ce de deux manières :

- en contrôlant de manière plus fine les productions contextuelles autour de la zone de recollement;
- en complexifiant le mode de calcul de la fonction de classement des solutions produites.

Nous n'exposerons ici que l'idée des extensions envisagées, dans la mesure où, dans l'ensemble, les améliorations des taux de justesse qu'elles ont permis d'obtenir sont restées relativement marginales.

Utilisation d'heuristiques

L'examen des erreurs révèle qu'un nombre important de prononciations incorrectes sont dues soit à l'insertion soit à la suppression d'un phonème. La cause de ces erreurs est simple à comprendre lorsque l'on examine de l'exemple suivant. Supposons que les mots *fil* et *collage* fassent partie du lexique de mots connus, et que le mot *filage* soit lui inconnu. La recombinaison des prononciations, telle que nous l'avons décrite, conduit à former la prononciation (incorrecte) /fiʒ/, dans lequel le *l* n'a pas de contrepartie phonétique. Ce qui s'est produit, c'est que notre algorithme a identifié un recouvrement graphique de taille 1 entre les deux lexies (le *l*, justement), et que cette lettre est alignée sur un phonème nul dans les deux prononciations utilisées. Le résultat du croisement, schématisé sur la figure 3.4, conserve l'appariement (*l*, /-/), ce qui se traduit par la suppression d'un phonème par rapport à la transcription correcte.

$$\begin{array}{rcl}
 & \text{collage} & \text{koll-aʒ-} \\
 + & \text{fils} & + \text{fi-s} \\
 \hline
 & \text{fillage} & \text{fi-aʒ-}
 \end{array}$$

FIG. 3.4 - Un exemple de recombinaison erronée.

Afin de supprimer ces erreurs, ainsi que d'autres du même acabit, nous avons ajouté des heuristiques de filtrage après le test de la ligne 10 dans l'algorithme de la page 91.

Heuristique 3.2.1 *Filtrer les solutions telles que la zone de recouvrement ne contient que des graphèmes alignés sur des nuls phonétiques.*

Cette première heuristique apporte une solution directe au problème que nous venons de mentionner. La seconde heuristique traite du même problème, mais s'applique aux prononciations construites par simple juxtaposition de deux segments, et qui, ayant un recouvrement de taille nulle, ne sont pas concernées par le premier filtre.

Heuristique 3.2.2 *Filtrer les solutions telles que :*

1. *le recouvrement graphémique soit de taille nulle;*
2. *le dernier caractère du segment initial ou le premier caractère du segment final soit aligné sur un nul phonémique.*

Ces heuristiques reflètent une connaissance *a priori* des difficultés que rencontre le mécanisme de recollement dans le cadre particulier de la transcription, à savoir que l'émission de nuls dans une chaîne phonétique est un phénomène hautement contextuel, et qu'à ce titre, il est plus prudent de restreindre, voire d'interdire leur présence autour de la zone de recollement. On notera que ces heuristiques sont désactivées si la procédure ne fournit aucune solution : dans ce cas, une dernière tentative est effectuée en supprimant ces filtres. Ceci permet de ne pas augmenter le taux de silence.

Utilisation de « bandages » *n*-grammatiques

La même idée est à l'œuvre dans une seconde variante de notre algorithme. La différence est que ces nouvelles heuristiques, qui fonctionnent sur le même principe que celles présentées précédemment, permettent d'acquérir les règles de filtrage de façon automatique.

Pour ce faire, pour une taille *n* fixée *a priori*, la phase de stockage du lexique inclut un apprentissage additionnel, qui consiste à enregistrer de toutes les séquences graphémiques de taille inférieure ou égale à *n* présentes dans le lexique, ainsi que leur(s) transcription(s) phonétique(s). Lors de la construction de nouvelles prononciations, on rejette les recombinaisons qui contiennent dans la zone de recollement des appariements entre séquences de taille *n* qui n'apparaissent pas dans notre dictionnaire de *n*-grammes.

Si l'on reprend l'exemple précédent, on s'aperçoit alors que l'enregistrement de toutes les séquences de taille 3 suffit pour invalider la prononciation /fi-aʒ-/ construite pour le mot *filage*. En effet, il n'existe pas de lexie en français telle que le segment *ila* soit mis en correspondance avec la séquence /i-a/. Cette idée, qui généralise la précédente, au prix d'une augmentation sensible du volume de données à stocker, a de fait été largement étendue et simplifiée dans un algorithme qui sera présenté à la section 3.3.

Fonctions de reclassement

Notre algorithme, dans de nombreux cas, et ce d'autant plus que le recouvrement est faible (et que, en conséquence, les contraintes de compatibilité associées sont plus lâches), construit des solutions multiples. Dans la version de base de l'algorithme, ces solutions sont classées, à taille de recouvrement égale, en fonction du nombre de paires de mots qui permettent de les construire. Cette façon de procéder a principalement trois effets pervers.

Il est tout d'abord important de remarquer que le test de la ligne 15 (page 91) provoque l'arrêt de la procédure à un niveau de recouvrement donné dès qu'une solution est exhibée. L'examen des erreurs suggère que ceci nous met à la merci de situations exceptionnelles, dans lesquelles un des mots servant de base à la transcription présente une prononciation très irrégulière.

L'alternative, consistant à considérer que toutes les solutions, indépendamment du recouvrement induit, sont uniformément valides, est lui aussi pénalisant. Ceci est d'autant plus vrai que le recouvrement est faible, et que le nombre de solutions proposées croît. Dans le cas limite où la taille du recouvrement est nulle, on se trouve même confronté à des situations où un segment particulièrement irrégulier est à la source d'un grand nombre de constructions, parce que le segment complémentaire est un suffixe ou un préfixe particulièrement fréquent.

Pour essayer de contourner ces problèmes, nous avons mis en œuvre différentes stratégies de reclassement des solutions. La première a consisté à supprimer le test de la ligne 15, de manière à construire toutes les solutions possibles, et ainsi s'affranchir du risque d'utiliser une lexie peu régulière dans nos solutions. Ce faisant, nous avons malheureusement contribué à renforcer le second effet : la majorité des solutions sont construites pour des faibles niveaux de recouvrement. Cette manière de procéder est également insatisfaisante, car elle conduit à remettre en cause un principe important de notre démarche, à savoir la préférence donnée à un critère de classement structurel (un long recouvrement) plutôt que fréquentiel. Là encore, diverses parades ont été envisagées.

Une première idée est de n'incrémenter le score d'une solution qu'à chaque fois que sa construction est due à des couples de segments différents, et non plus à des couples de mots différents. Dans le même ordre d'idée, nous avons également essayé de favoriser les solutions qui étaient proposées dans le plus grand nombre de contextes (graphiques) différents. Nous avons également envisagé d'intégrer dans la fonction de reclassement la fréquence des « bandages n -grammatiques », comme une mesure complémentaire de la qualité du recollement. Nous n'entrons pas plus avant dans le détail exact de toutes ces tentatives, dans la mesure où, répétons-le, les performances globales n'ont été que marginalement modifiées par ces extensions.

3.2.5 Remarques finales

À propos des schémas de construction

Un des intérêts de la stratégie développée dans JUPA est qu'elle fonde la construction des transcriptions d'un mot inconnu sur un nombre relativement

petit de lexies, nombre qui d'ailleurs est d'autant plus faible que le recouvrement, et donc le niveau de confiance que l'on peut accorder à la solution produite est plus grand. Ceci permet notamment de donner un sens plus précis à la notion de mot analogue, (« est analogue tout mot qui contribue à la réalisation de la prononciation finale », par exemple) et d'examiner pour chaque solution les exemples qui ont participé à la formation du mot. Cet examen (principalement conduit sur les données des lexiques NETTALK et ENST) permet de mieux saisir les motifs du succès (relatif) de cette approche, ainsi que d'identifier quelques causes d'erreur supplémentaires. On peut en effet grossièrement classer les solutions en quatre familles principales.

Les paires minimales Dans une première classe, on trouve les solutions qui s'appuient sur l'utilisation d'un mot du lexique avec lequel le mot cible fonctionne comme une paire minimale. Dans ce schéma, un des deux mots utilisés (le second dans les exemples que nous présentons) contribue beaucoup plus de à la solution que l'autre. Comme exemple de ce type de construction, citons *waffle* = *wacky* + *baffle*; *staple* = *stable* + *maple*; ou encore *sprawl* = *sprain* + *brawl*. On peut considérer que ce schéma fournit dans l'ensemble de bons analogues, la rime en particulier jouant un rôle prépondérant, et des solutions plutôt correctes.

Les pseudo-paradigmes Un second type de schémas particulièrement bien attesté se fonde sur l'utilisation d'un long segment initial, appartenant à un mot relié sémantiquement au mot à transcrire. Le second terme (final) fournit quand à lui, dans les cas favorables, la séquence correspondant à une dérivation morphologique afférente. Quelques exemples de ce type de schémas permettent d'en apprécier la productivité; ainsi *anglicization* = *anglicize* + *hybridization*; *goodly* = *good* + *allegedly*; ou encore (notre préféré) *morphological* = *morphologist* + *analogical*. On observe également, de façon plus marginale, le même phénomène mettant en jeu des préfixes. Ces constructions-là posent néanmoins des problèmes à notre algorithme, et sont en fait cause de nombreuses erreurs. Les raisons sont diverses, citons-en trois principales :

- La ressemblance entre les deux segments initiaux est fortuite, et on n'a pas en fait affaire à un terme sémantiquement proche. Ainsi la construction de la prononciation du mot *albumen* sur la base du mot *album* (adjoint à *abdomen*), ou encore de *rapist*, incorrectement dérivé de *rapid* et de *philanthropist*.
- La ressemblance est réelle mais la dérivation morphologique s'accompagne d'un changement de qualité d'une ou de plusieurs voyelles. Ces alternances ne peuvent être prises en compte dans notre mécanisme, qui, privilégiant les longs recouvrement, essaie d'utiliser « le plus possible » du segment initial. Par exemple, la prononciation de *abjuration*, improprement dérivée de la recombinaison de *abjure* et de *declaration*, car manquant en quelque sorte l'alternance /ur/ -/ərefən/ corrélative à la dérivation. Les constructions de ce type-là sont de fait relativement hasardeuses, et dépendent pour leur succès de la disponibilité d'un segment final excédant suffisamment le suffixe.

- Un cas inverse est également attesté, dans lequel c'est le terme dérivé qui est mis à contribution pour inférer la prononciation du terme plus simple. Un tel exemple a été longuement discuté dans (Federici, Pirrelli, & Yvon 1995), celui du couple *Martin* - *Martine*. Le problème ici, c'est que l'utilisation du mot le plus long pour prononcer le plus court nous met à la merci d'une seule exception, qui viendra masquer la véritable alternance morphologique. Dans cet exemple, la présence d'un mot « exceptionnel » tel qu'*Amin* (lire /amin/) dans notre lexique, qui s'accorde avec *Martine* sur la prononciation de la partie finale, suffit à entraîner la construction de la transcription /martin/ pour *Martin*. En effet, la prévalence dans l'élaboration des solutions de longs recouvrements des segments *phonétiques* entraîne le choix de cette solution, fondée sur un recouvrement phonétique de taille 2 (/in/). Elle sera en conséquence préférée à n'importe quelle solution alternative qui refléterait la réelle alternance morphologique, et qui, dans ce cas précis, se fonderait sur un recouvrement de taille nulle (par exemple la recombinaison de *Martine* avec *Constantin*). Nous reviendrons plus en détail sur cette erreur troublante à la section 4.1.2.

Ce qu'il est important de noter ici, c'est que cette stratégie de recollement est de fait aveugle à l'existence, même massivement (statistiquement) attestée, de paradigmes morphologiques. Quand malgré tout certaines solutions correctes sont fabriquées qui tirent profit la structure particulière de ces relations lexicales, cela est principalement dû à des circonstances fortuites.

Remarquons enfin que le succès de JUPA est ici une conséquence directe de la forme spécifique de la morphologie des langues que nous avons étudiées, à savoir l'importance (en productivité) des paradigmes dérivationnels. Il est probable que les constructions de ce type contribueraient de manière moins sensible aux résultats si nous avions travaillé sur des données turques ou hongroises. La morphologie plus strictement concaténative de ces langues se laisserait en effet moins facilement saisir au travers de constructions ne mettant en jeu que *deux* segments de mots.

Les mots composés Une propriété caractéristique de l'anglais est à l'origine d'une autre grande classe de constructions, à savoir celles qui correspondent à simple juxtaposition de deux segments d'exemples disponibles. Le schéma de construction de nouveaux éléments lexicaux par simple concaténation de mots existants est en effet très productif en anglais, et est attesté sous diverses formes (N+N, A+N, V+N, etc) (Spencer 1991). Ce schéma a de plus la vertu, au regard du mécanisme ici évalué, de ne pas entraîner de bouleversement majeur au niveau des composants phonologiques des éléments composés, puisque seule la distribution des accents est affectée. Il est en effet couramment admis (Chomsky & Halle 1968) que cette composition redistribue l'accent principal sur le premier élément du composé. Ainsi, le composé : *blackbird*, formé par concaténation d'un adjectif : *black* et d'un nom : *bird*, reçoit un accent lexical principal sur sa première syllabe; la seconde syllabe est désaccentuée et ne porte qu'un accent secondaire. On notera, *a contrario*, que ce fait compromet le succès des solutions produites par simple juxtaposition lorsqu'on soumet JUPA à la tâche de prédiction des positions accentuées sur un lexique anglais (voir la table 3.5).

Corpus	Correction		Justesse	
	% mots	% symboles	% mots	% symboles
NETSTRESS	51,100	84,448	54,668	92,168
BDLEX	78,734	86,933	86,080	97,914
CMU-20	40,933	47,752	73,244	89,906
ENST-20	78,368	80,988	91,904	95,954
ILC-20	85,832	92,795	89,263	98,943
KUN	83,616	88,505	91,434	98,180
TUB	32,383	51,068	51,388	91,416

TAB. 3.5 - Performances de JUPA sur différents lexiques.

La productivité morphologique de la juxtaposition contribue largement au fait que de nombreuses transcriptions sont directement fabriquées par concaténation des transcriptions de deux mots disponibles, ou qu'à l'inverse, un mot entier soit trouvé intégralement dans un composé plus long. La première situation est exemplifiée par le cas de *hangdog*, construit sur la base de *hand* et de *dog*; ou encore de *saucepan*, sur la base de *sauce* et de *pan*. Un exemple de la seconde est *walk*, transcrit par analogie avec *wall* et *catwalk*. Une configuration voisine est également très productive : celui où l'une des parties du mot composé est disponible dans un autre mot composé, comme par exemple dans la construction de la prononciation de *fanlight*, à partir de *fan* et de *sunlight*. L'existence de composés de ce type est pour beaucoup dans le taux de correction non-négligeable des solutions construites sans aucun recouvrement (voir la figure 3.2 de la page 93).

Les constructions hasardeuses Le quatrième groupe de constructions rassemble toutes les autres, qui se laissent moins bien classer, et que l'on doit uniquement au hasard. Ce groupe reste malgré tout le plus nombreux, et les exemples abondent. Citons-en quelques-uns, *baggy* construit par recombinaison de *baggage* et de *groggy*; *oasis* par recombinaison de *oaf* et de *basis*; ou encore *morgue* par recombinaison de *moral* et de *argue*.

Derniers résultats et perspectives

Pour compléter notre description des performances de cette stratégie de recollement, nous donnons dans la table 3.5 les résultats des évaluations conduites avec les autres lexiques dont nous disposons.

Si ces résultats varient fortement en fonction de la difficulté des lexiques, on notera que cette la stratégie de transcription mise en œuvre à travers JUPA donne toujours, en termes de justesse par mot, des résultats équivalents ou meilleurs que ceux de DEC (voir la table B.6, en page 234). En ce qui concerne les taux de silence, on observe également une très forte dispersion, allant de 3,3 % pour le lexique ENST-20, à près de 33 % pour le lexique allemand.

En conclusion, les expériences menées avec cet algorithme simpliste se sont avérées dans l'ensemble suffisamment concluantes pour que plusieurs extensions soient envisagées. La fin de ce chapitre sera partiellement consacrée à la présen-

tation d'une généralisation de cette procédure (voir en particulier la section 3.3), généralisation qui autorise la prise en compte des segments internes des mots du lexique.

En anticipant un peu, notons qu'un autre type de généralisation sera introduit dans le chapitre suivant (pages 137 et suivantes), qui visera autant à compenser quelques-unes des lacunes de ce système, qu'à fonder plus spécifiquement notre approche sur l'hypothèse que la présence de paradigmes morphologiques (flexionnels et dérivationnels) contribue à structurer en profondeur le lexique; et que la forme particulière que ces paradigmes imposent aux données linguistiques permet de mettre en œuvre avec succès des procédures combinatoires spécifiques s'appuyant explicitement sur ces structures. En d'autres termes, c'est une tentative de spécialisation de JUPA que nous présenterons à cette occasion; spécialisation qui prendra la forme de procédures *explicitement* conçues pour construire des analogies se fondant sur des alternances paradigmatiques.

3.3 Transcription par recombinaison n-aire de segments

Nous avons présenté à la section précédente une première approche de la prononciation par analogie, ainsi que sa mise en œuvre dans l'algorithme JUPA, et son évaluation sur divers corpus. Une approche complémentaire de ce type de mécanisme a été élaborée par Dedina et Nusbaum (Dedina & Nusbaum 1991), et est présentée en détail dans l'annexe C. Ce que cette démarche alternative nous a principalement révélé, c'est qu'il existe un moyen simple de se débarrasser du problème du silence, qui consiste à utiliser non plus deux segments (un segment initial et un segment final), mais à mettre à contribution toutes les sous-chaînes d'un mot cible disponibles dans le lexique d'exemples.

Dans cette section, nous développons une extension naturelle de ces deux approches, qui étend à la recombinaison d'un nombre quelconque de segments la démarche mise en œuvre dans JUPA. En particulier, le nouvel algorithme que nous proposons persévère dans le principe, à nos yeux important, de la sélection des transcriptions concurrentes d'un mot cible sur les bases d'un critère structurel, plutôt que fréquentiel.

Nous présenterons dans un premier temps les fondements de cette extension, en reformulant notre programme comme la recherche d'un sous-ensemble de recouvrement phonémique maximal de la chaîne cible. Dans un second temps, nous formaliserons cette démarche dans un algorithme, SMPA. Nous montrerons ensuite, sur la base de résultats numériques, que cette approche permet d'obtenir des performances sensiblement supérieures aux approches alternatives présentées dans les sections et chapitres précédents. Nous aborderons finalement les limites de cette démarche, principalement en relation avec la conception plus formelle de l'analogie que nous cherchons à défendre dans ce document.

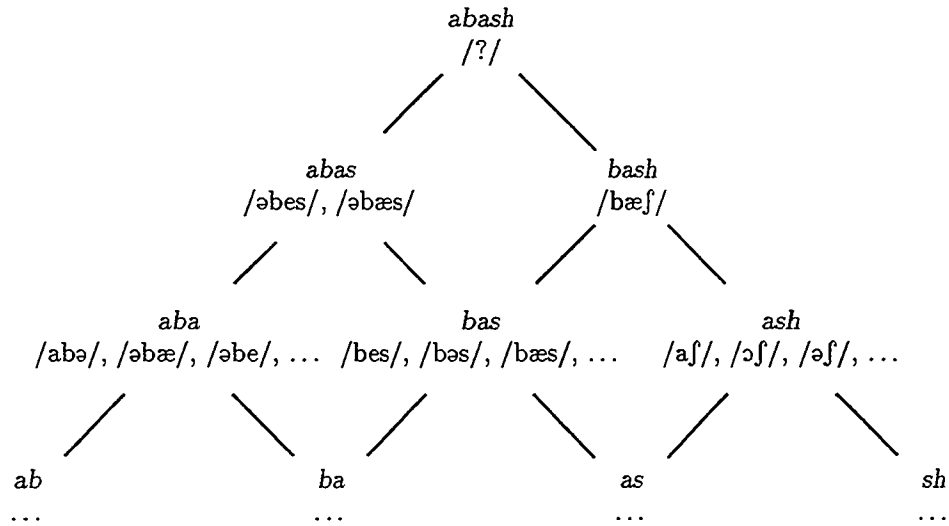
3.3.1 Quelques concepts

Avant d'entrer dans les détails de la présentation formelle, essayons de préciser l'intuition qui nous a guidée pour introduire cet algorithme. Nos expériences avec JUPA nous ont convaincu de la pertinence de la remarque suivante : si l'on cherche à prononcer le mot inconnu $x = c_1 \dots c_{l(x)}$, le mieux que l'on puisse faire, c'est de recombinaison les prononciations des sous-chaînes $c_1 \dots c_{l(x)-1}$ et $c_2 \dots c_{l(x)}$, pour autant que ces prononciations soient connues, et que la recombinaison soit possible. Sinon, il va falloir se contenter de chaînes plus petites, tout d'abord en essayant de recombinaison $c_1 \dots c_{l(x)-2}$ et $c_2 \dots c_{l(x)}$, ou encore $c_1 \dots c_{l(x)-1}$ et $c_2 \dots c_{l(x)-1}$, ou encore $c_1 \dots c_{l(x)-2}$, $c_2 \dots c_{l(x)-1}$ et $c_3 \dots c_{l(x)}$, etc. L'idée est dans tous les cas de trouver dans notre lexique (considéré comme un stock de segments) un ensemble de chaînes, qui d'une part recouvrent complètement le mot à transcrire, et qui d'autre part sont en correspondance avec des transcriptions qui puissent être recombinaisonnées (nous verrons plus loin quel sens précis à donner à cette contrainte de « recombinaisonabilité »).

Une autre façon de saisir cette intuition est de remarquer que si la prononciation d'un mot x est connue, alors toutes ses sous-chaînes sont disponibles dans le lexique, et avec des prononciations qui sont recombinaisonnables. Pour un mot cible donné, qui ne vérifie donc pas cette propriété, nous considérerons que la meilleure transcription, c'est celle qui permet de se rapprocher au mieux de cette situation idéale. Pour cela, nous allons rechercher dans le stock de segments disponibles *le plus grand sous-ensemble de segments recombinaisonnables du mot cible*. Ces segments peuvent être vus comme des *traces* ou des indices de la prononciation correcte. La procédure que notre algorithme met en place peut alors être considérée comme la recherche de la chaîne phonétique dont « le plus » de sous-chaînes sont attestées. À titre d'exemple, considérons le mot inconnu *abash*, et cherchons toutes les sous-chaînes (de taille supérieure à 2) de ce mot dans le lexique NETTALK. Cet ensemble, ainsi que les diverses prononciations afférentes, est partiellement reproduit sur la figure 3.5.

Il est possible, à partir de cet exemple, d'illustrer le concept de « recombinaisonabilité » introduit dans les paragraphes précédents. Par exemple, bien que deux segments graphiques de taille 4 (*abas* et *bash*) recouvrent complètement le mot cible, leurs prononciations ne sont pas *recombinaisonnables*. Considérons en effet les prononciations /əbes/ (correspondant au premier segment) et /bæf/ (correspondant au second) : dans la mesure où le premier segment phonétique se termine par une séquence (/es/) qui n'apparaît pas dans le second, l'aboutement de ces deux segments est impossible.

Pour élaborer une prononciation du mot *abash*, il faudra donc se fonder sur un recouvrement moins large de la chaîne graphique. Par exemple en utilisant le segment initial *aba*, et en le recombinaisonnant avec le segment final *bash* : ces deux segments-ci sont effectivement compatibles, puisque l'on peut abouter deux des chaînes phonétiques correspondantes, à savoir /əbæ/ d'une part et /bæf/ d'autre part. Il existe naturellement d'autres solutions, ne serait-ce qu'en utilisant les segments *aba* et *ash*, pour lesquels diverses recombinaisonnements sont permises (ainsi /abə/ et /əf/, repérable sur le graphe de la figure 3.5). Toutefois, aucune autre solution n'exhibe de recouvrement phonétique de taille supérieure à 1. C'est donc la solution /abæf/ qui doit être préférée.

Figure 3.5: Extraction des sous-chaînes du mot *abash*

Bien que l'exemple précédent décrive la construction d'une transcription par recollement de deux segments, le principe mis en œuvre est en fait plus général que celui de JUPA, parce que le meilleur recouvrement est recherché dans l'ensemble de toutes les recombinaisons licites, et non plus simplement sur les recombinaisons licites de taille 2.

3.3.2 Description formelle de l'algorithme

Nous présentons dans cette section la mise en œuvre algorithmique du principe de recherche du plus grand ensemble de sous-chaînes compatibles énoncé dans les paragraphes précédents. Cet algorithme se déroule en trois étapes principales, que nous étudierons tout à tour, qui sont l'étape d'extraction et de stockage des segments, la construction du graphe de transcriptions, et la construction et le classement des prononciations.

Stockage du lexique

Dans la mesure où la procédure mise en œuvre fait un appel extensif à l'appariement de segments avec des segments de mots du lexique, nous avons choisi de ne pas stocker le lexique en l'état. La phase « d'apprentissage » de notre algorithme consiste donc à extraire tous les segments possibles des mots connus, pour les stocker dans un arbre lexical. C'est le propos de la procédure suivante (algorithme 2).

L'utilisation d'un arbre lexical pour stocker les segments permet d'éviter l'examen complet du lexique à chaque mot nouveau. Une chaîne inconnue x étant donnée à transcrire, la construction du graphe de prononciation ne requiert

Algorithme 2: Stockage du lexique

Entrée: Un lexique de prononciations $\mathcal{L}$

Sortie: Un arbre lexical contenant toutes les sous-chaînes d'un mot du lexique, ses prononciations, et les fréquences associées

```

(1)   $T$  est un arbre lexical vide.
(2)  pour tout  $x = c_1 \dots c_{l(x)} \in \mathcal{L}$ 
(3)    pour tout  $z \in \mathcal{L}(x)$   $l(z) \geq 2$ 
(4)      si  $z$  existe dans  $T$  alors incrémenter sa fréquence
(5)      sinon insérer  $z$  dans  $T$ 
(6)    fin
(7)  fin

```

plus que $|\mathcal{L}(x)|$ recherches linéaires dans T et évite ainsi le parcours intégral du lexique. Cette façon de procéder nous a également permis d'introduire deux nouveaux paramètres, qui sont les tailles minimum et maximum d'un segment stocké dans T .

Concernant l'utilisation d'alignements comportant des nuls graphiques, il en va dans cette procédure comme pour JUPA (voir la discussion de la section 3.2.2). En effet, la seule information nécessaire durant l'étape de stockage est la connaissance des alignements entre segments graphiques et phonétiques. Modulo la fixation d'une convention de réaffectation des phonèmes correspondants, il est possible de se débarrasser de ces nuls graphiques, et d'ignorer leur existence dans le reste de l'algorithme.

Construction des solutions

Le cœur de notre algorithme est la construction d'un graphe de prononciation $\mathcal{G}$, dont les caractéristiques sont les suivantes :

- $\mathcal{G}$ contient deux sommets S (start) et E (end), qui ont des ensembles d'arêtes entrantes (respectivement sortantes) vides.
- Les autres sommets de $\mathcal{G}$ correspondent à des couples $S = (t, i)$, où t est une séquence de phonèmes $t = p_1 \dots p_k$, telle qu'il existe dans le lexique un mot M contenant une sous-chaîne $z = c_1 \dots c_k$ du mot inconnu x , et z est alignée avec t dans la transcription de M ; et i est l'indice dans x du début de z . À chaque sommet est associé la fréquence absolue d'apparition dans le lexique du couple de séquences. Cette pondération sera utilisée, nous le verrons, dans la procédure de classements des solutions.
- $\mathcal{G}$ contient une arête de $S_a = (t_a, i_a)$ vers $S_b = (t_b, i_b)$ si et seulement si $k = i_b - i_a > 0$ et $k \leq l(t_a)$; et si $\exists (k', k'')$ positifs tels que : $p_{a,k'} \dots p_{a,l(t_a)} = p_{b,1} \dots p_{b,k''}$. En d'autres termes, il y a une arête entre deux nœuds si les chaînes phonétiques correspondantes se chevauchent, à la manière des comptines pour enfant « J'en ai marre - Marabout - Bout d'ficelle », etc. Ces arêtes sont valuées par k'' , qui représente la taille du recouvrement entre les deux chaînes t_a et t_b .
- $\mathcal{G}$ est complété par en reliant S à tous les sommets d'indicateur de position égal à 1, et E à tous les sommets d'indicateur de position égal à $l(x)$. Ces

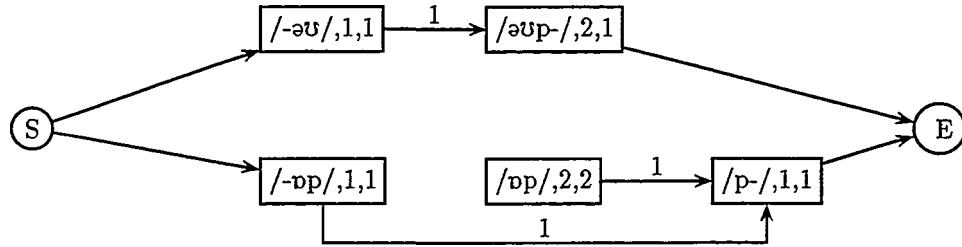


Figure 3.6: Construction du graphe d'analogies

arêtes ont une valuation nulle.

Considérons le lexique constitué des quatre entrées suivantes :

- (*photo*, /f-əʊtəʊ/);
- (*shop*, /ʃ-ɒp/);
- (*slop*, /slɒp/);
- (*slope*, /sləʊp-/).

Supposons de plus que le mot cible soit *hope*. Tous les segments (à l'exception du mot complet) qui composent le mot inconnu apparaissent au moins une fois dans le lexique : ainsi *hop* dans *shop*, *ope* dans *slope*,... Le graphe $\mathcal{G}$ correspondant à l'examen de ce lexique pour construire la prononciation du mot *hope* est reproduit sur la figure 3.6. Sur ce graphe, on a fait figurer, pour chaque sommet, la chaîne phonétique associée, l'indice de début, et la fréquence de l'association entre les sous-chaînes graphiques et phonétiques dans le lexique. Ainsi, le nœud étiqueté par /ɒp/ correspond à une sous-chaîne de *hope* commençant en position 2 (*op*), et cette association apparaît deux fois dans notre lexique (dans *slop* et dans *hop*).

Une fois $\mathcal{G}$ construit, chaque chemin dans le graphe permettant de joindre les nœuds initial et final correspond par construction à une prononciation possible du mot inconnu. Ainsi, les deux chemins de *S* à *E* dans le graphe représenté sur la figure 3.6 permettent de construire les deux prononciations /-ɒp-/ et /-əʊp-/. On notera toutefois que ces deux prononciations sont incorrectes, ce qui n'est pas surprenant puisque les exemples dont on s'est inspiré pour prononcer le *h* sont deux mots dans lesquels ce graphème fonctionne au sein d'un digraphe (*sh* dans *shop* et *ph* dans *photo*).

Ordonnancement des solutions

Une fois le mécanisme de production des transcriptions donné, il reste à construire un critère non fréquentiel qui permette, conformément à nos intentions, de classer ces transcriptions. Pour ce faire, considérons la fonction qui associe à chaque chemin de *S* vers *E* (i.e. à chaque transcription) la moyenne de

la longueur des segments attachés aux nœuds internes (S et E exclus) traversés. Compte-tenu de la procédure de construction de notre graphe, il est clair que pour un mot cible x donné, cette fonction est bornée supérieurement par $l(x)$ ³. Cette valeur est atteinte si et seulement si la forme graphique à transcrire se trouve déjà dans le lexique (éventuellement en tant que sous-chaîne d'une autre forme⁴), auquel cas il existe un segment phonémique entièrement apparié à x (de taille $l(x)$), et un chemin dans le graphe qui ne traverse que le nœud correspondant à ce segment. Si la forme cible n'existe pas intégralement dans le lexique, alors la valeur maximale de cette fonction est $l(x) - 1$, correspondant au cas où il existe une solution produite par recombinaison de deux segments, etc.

Pour en dériver de cette fonction un critère permettant de comparer la prononciations de mots cibles de taille différente, il suffit ensuite de ramener cette moyenne à la taille du mot cible. Le score sur la base duquel les chemins dans $\mathcal{G}$ (i.e. les transcriptions) sont ordonnés est alors, en notant P un chemin, s un segment associé à un nœud quelconque de ce chemin, et $|P|$ le nombre de nœuds internes :

$$C(P) = \frac{\sum_{s \in P} l(s)}{|P| \times l(x)} \quad (3.1)$$

La figure 3.7 reproduit les valeurs possibles de ce score en fonction de la taille du mot cible. Sur cette figure sont également reportés les meilleurs scores possibles pour un nombre de segments donnés. En se reportant à ce graphique, on note ainsi que pour un mot de 6 lettres, le meilleur score possible d'une solution construite en recombinant trois segments est légèrement inférieur à 0,7.

L'examen de la figure 3.7 permet de mettre en évidence la propriété importante qu'exhibe ce critère de classement, à savoir le compromis qu'il permet d'effectuer entre le nombre de segments utilisés, et la taille moyenne des recouvrements. Ceci se lit clairement par exemple pour les mots de longueur 8 : le quatrième meilleur score correspond à une recombinaison de trois segments, combinaison qui par conséquent est meilleure que la plupart des combinaisons de deux segments.

Dans le cas où deux chemins ont le même score, ils sont secondairement départagés en raison de la fréquence cumulée des nœuds qu'ils traversent. De nombreux modes de calcul alternatifs ont été évalués, avec pour objectif de prendre en compte, dans un score unique, le critère de recouvrement maximal et le critère fréquentiel. Les meilleurs performances ont néanmoins été obtenues en utilisant une procédure de classement en deux temps.

Si l'on reprend l'exemple du mot cible *hope*, les deux transcriptions

3. En toute rigueur, ceci est inexact : l'utilisation d'alignements qui contiennent des nuls graphiques autorise que certaines formes graphiques du lexique soient de longueur inférieure à la chaîne phonétique correspondante, et donc, que des valeurs supérieures à $l(x)$ soient observées. Pour la simplicité de l'exposé, nous ignorerons pour l'instant ces configurations un peu atypiques. Leur existence ne remet pas en cause la validité de notre critère de classement.

4. Dans la pratique, nous nous sommes débarrassés de ce cas-là, qui conduit de nombreuses aberrations. Le cas de *Martin* prononcé /martin/ par analogie avec *Martine* est un exemple typique. Pour ce faire, nous avons introduit des graphèmes spéciaux de début et de fin de mot. Dans ce contexte, alors la valeur $l(x)$ ne peut être atteinte que dans le cas où la forme cible est exactement retrouvée dans le lexique.

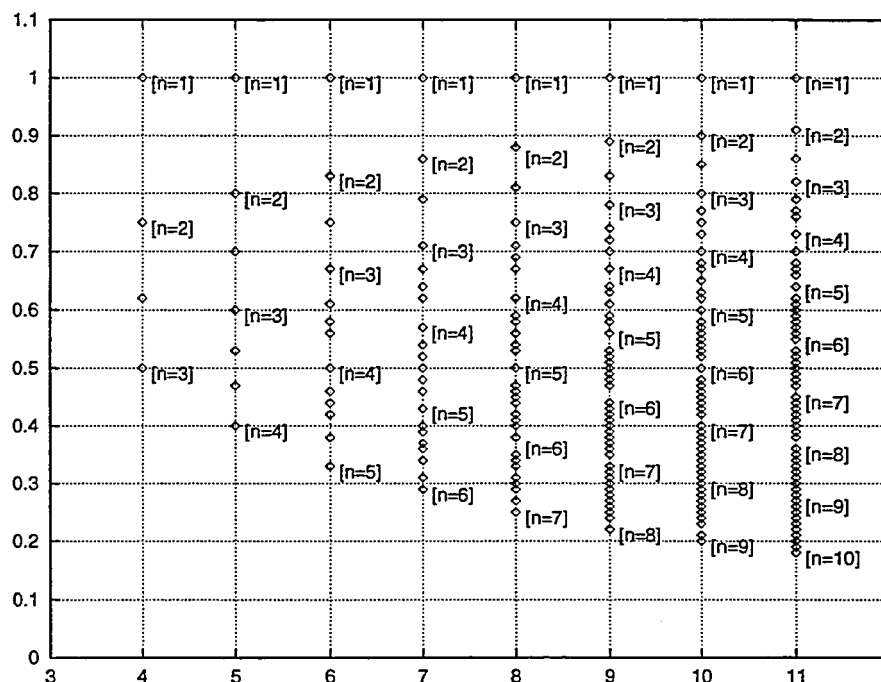


FIG. 3.7 - Critère principal de classement des solutions de SMPA

(/vp/ et /əv/) auxquelles nous sommes parvenu ne peuvent en définitive être départagées : leur score (critère du recouvrement maximal) est de $(3 + 2)/(2 \times 4) = 0,625$, et les segments qu'elles utilisent correspondent tous à des associations de fréquence unité.

L'algorithme complet

Une présentation synthétique de notre algorithme (baptisé SMPA) est donnée ci-dessous (algorithme 3).

Algorithme 3: Prononciation par recombinaison de n segments - SMPA

Entrée: Un lexique $\mathcal{L}$, une chaîne $x = c_1 \dots c_{l(x)}$ à transcrire

Sortie: Un ensemble de transcriptions possibles trié de la plus à la moins plausible

SMPA(x)

- (1) $\phi(x) \leftarrow \lambda$
- (2) Construire itérativement par appariement avec tous les mots du lexique le graphe $\mathcal{G}$
- (3) Construire tous les chemins de S à E dans $\mathcal{G}$ et les classer par ordre de score croissant et de poids (calculé comme la somme des fréquences des sommets) croissant.
- (4) Proposer les solutions correspondantes par score décroissant.

Concernant la complexité de cet algorithme, il est important de remarquer

que la construction de toutes les transcriptions n'est pas nécessaire. La procédure de recherche des sous-chaînes dans le lexique, menée de la gauche vers la droite, permet d'associer, dès l'étape de construction de $\mathcal{G}$, un numéro topologique à chaque sommet. Dans la mesure où ce graphe est par définition sans cycle, le plus court chemin est obtenu au prix d'une seule exploration de chacun des nœuds. Comme il y a exactement $l(x) \times (l(x) - 1)$ sous-chaînes graphiques de taille supérieure ou égale à 2 d'une chaîne x ; et en supposant l'ambiguïté phonémique d'une chaîne graphique quelconque bornée supérieurement, la procédure SMPA produit des transcriptions en un temps $O(l^2)$ par rapport à la taille de la cible.

Ce nouvel algorithme est une généralisation de JUPA, puisqu'il permet de tirer parti de la connaissance de la prononciation de sous-chaînes intérieures du mot inconnu. Dans les cas où le lexique d'exemples permet de construire une solution par recombinaison de deux sous-chaînes du mot cible avec un large recouvrement, notre algorithme donne exactement les mêmes résultats que JUPA, et c'est heureux, puisque les configurations à large recouvrement correspondent le plus souvent à des solutions correctes (voir la discussion de la page 93). Par contre, dans les cas où le recouvrement est trop faible, comme dans ceux pour lesquels il n'existe pas de prononciation qui n'utilise que deux segments, notre algorithme permet, en proposant des solutions fondées sur un plus grand nombre de « morceaux », de trouver des transcriptions, éventuellement meilleures que l'algorithme précédent. On aura l'occasion de vérifier numériquement cette assertion à la section 3.3.3.

On notera qu'il généralise également le programme de Dedina et Nusbaum (voir l'annexe C), dans la mesure où la création d'une arête dans le graphe des prononciations n'est plus simplement conditionnée par l'existence d'un phonème commun, mais par un nombre *quelconque* de phonèmes en commun. Ce faisant, il permet à nouveau d'utiliser un critère de recouvrement maximal entre les segments du lexique utilisés pendant la phase de recombinaison.

3.3.3 Performances

Nous présentons dans cette section les résultats d'une évaluation complète de l'algorithme SMPA. Nous examinerons successivement l'influence du facteur de compression du lexique, évalué par les contraintes de taille des segments recombinaison, puis les performances sur différents lexiques de taille comparable. Enfin, nous comparerons les performances de cette approche avec l'algorithme classificatoire de référence, DEC, et avec l'approche analogique plus simple mise en œuvre dans JUPA, en discutant les résultats obtenus sur le lexique NETTALK et le comportement asymptotique de ces trois algorithmes.

L'influence de la taille des segments

Le premier résultat important qui se lit sur la table 3.6 est que le passage d'une stratégie de recollement de deux segments à une stratégie permettant d'en utiliser un nombre plus grand permet effectivement de réduire de façon drastique le taux de silence résiduel. Ce taux de silence passe en effet, pour le lexique NETTALK, d'environ 6 % à un peu plus de 0,4 %. Une réduction très

Contrainte	% mots	% silence	% symboles
$2 \leq l \leq 8$	64.055	0.425	93.223
$2 \leq l \leq 6$	64.030	0.425	93.255
$2 \leq l$	63.965	0.425	93.192
$2 \leq l \leq 7$	63.820	0.425	93.202
$3 \leq l \leq 8$	63.420	2.320	91.756
$3 \leq l \leq 6$	63.315	2.320	91.767
$3 \leq l \leq 7$	63.140	2.320	91.723
$2 \leq l \leq 5$	61.411	0.444	92.730
$3 \leq l \leq 5$	60.656	2.350	91.219
$2 \leq l \leq 4$	51.244	0.469	90.236
$3 \leq l \leq 4$	50.531	2.350	88.795
$2 \leq l \leq 3$	22.950	0.400	81.596
$3 \leq l \leq 3$	22.767	2.150	80.318

TAB. 3.6 - Performances de SMPA en fonction de la taille des segments. Les données sont celles de NETTALK

sensible du silence est observée sur tous les lexiques que nous avons traité. On notera que pour le lexique allemand en particulier, le taux de silence résiduel reste relativement important (de l'ordre de 10 %). Compte-tenu toutefois de la très faible valeur du silence sur le lexique NETTALK, nous confondrons dans la suite de la présentation des performances de SMPA sur ce lexique les notions de correction et de justesse.

Un second point important mérite d'être signalé : la meilleure stratégie n'est pas celle qui consiste à conserver tous les segments disponibles dans le lexique d'exemples. La table 3.6, qui liste les résultats obtenus en termes de mots corrects, de silence, et de phonèmes corrects en utilisant diverses restrictions sur la taille l des fragments qui seront utilisés pour la construction de nouvelles prononciations, témoigne de ce phénomène.

Les meilleurs résultats sont en effet obtenus en éliminant du stock de segments ceux dont la longueur est supérieure à 8 (ou à 6, les résultats sont très comparables). Ceci signifie concrètement que stocker tout le lexique n'est pas nécessaire, mais que l'on peut en quelque sorte compresser l'information pertinente en ne considérant que des segments de taille de taille bornée. Un comportement analogue a été mis en évidence pour l'algorithme PRONOUNCE (voir l'annexe C).

Évaluation multilingue

Nous présentons dans cette section les résultats des évaluations de SMPA sur différents lexiques. Pour chacun d'entre eux, la table 3.7 fournit les performances de transcription obtenus pour la configuration la plus favorable des contraintes de taille.

Nous l'avons déjà signalé, le taux de silence résiduel sur le lexique allemand reste non-négligeable, voisin de 10 %. Les résultats reportés dans le tableau 3.7 correspondent à une justesse par mot de 57,419 %, et à une justesse par phonème

Corpus	Contrainte	% mots	% symboles
NETSTRESS	$2 \leq x \leq 258$	53,833	91,760
BDLEX	$2 \leq x \leq 6$	86,542	95,433
CMU-20	$2 \leq x \leq 8$	44,330	88,022
ILC-20	$2 \leq x \leq 7$	95,735	99,061
ENST-20	$2 \leq x \leq 6$	79,145	95,626
TUB	$2 \leq x \leq 8$	51,375	82,151
KUN	$2 \leq x \leq 258$	89,859	98,570

TAB. 3.7 - Performances de SMPA sur différents lexiques.

de 92,850 %. À l'exception de ce cas particulier, le silence est remarquablement faible, puisque le plus fort taux de silence (2,3 %) est observé sur le lexique BDLEX, et est largement imputable à la proportion non-négligeable de mots composés⁵ que ce lexique contient.

Pour le reste, la table 3.7 met de nouveau en évidence les différences importantes de difficulté entre les lexiques dont nous disposons. La comparaison de ces résultats avec les évaluations correspondantes de DEC (page 234), JUPA (page 103) et PRONOUNCE (page 241) permet de réaliser que la stratégie de recombinaison n-aire de segments est globalement la plus performante de celles que nous avons rencontrées jusqu'ici.

Comparaison avec DEC et JUPA

Nous avons cherché à comparer de manière plus fine les performances de SMPA avec notre algorithme de référence, DEC. La comparaison a été effectuée en utilisant pour chacun des algorithmes le jeu de paramètres le plus favorable : le contexte utilisé pour DEC s'étend selon le schéma *fnb*, alors que pour SMPA, on utilise uniquement des segments dont la taille est comprise entre 2 et 8 caractères. Les résultats sont résumés dans la table 3.8.

Comme on le constate, ces deux algorithmes rencontrent sensiblement les mêmes difficultés : parmi les mots que DEC échoue à transcrire correctement, un peu moins d'un tiers sont prononcés correctement par SMPA ; alors que seules 16,21 % des erreurs de SMPA sont corrigées par DEC. L'approche à base d'analogies se révèle néanmoins significativement supérieure à l'approche par construction d'arbre de décision : elle permet de gagner près de 8 points en termes de correction par mot sur ce lexique (la différence entre les résultats moyens est significative au seuil de 0,01).

La comparaison entre les deux stratégies de recombinaison de segments présentées dans ce chapitre est également instructive. Elle permet en particulier de mettre numériquement (voir la table 3.9) en évidence la forte parenté qui unit ces deux approches.

5. Les mots composés, transcrits par notre procédure comme des entités « monolithiques », sont mécaniquement responsables d'un grand nombre de non-transcriptions : pour pouvoir transcrire un mot composé de type *mot1-mot2*, il faut en effet trouver une sous-chaine du lexique qui contienne à la fois la partie finale de *mot1*, le tiret, et la partie initiale de *mot2*. Décomposer ces entrées lexicales en deux termes transcrits successivement permettrait sans

		SMPA		
		Correct	Faux	Total
DEC	Correct	89,72	10,28	100,00
		50,84	5,83	56,67
		79,38	16,21	
	Faux	30,49	69,51	100,00
		13,21	30,12	43,33
		20,62	83,79	
Total	64,06	35,95	100,00	
	100,00	100,00		

TAB. 3.8 - *Matrice de confusion entre les transcriptions de DEC et SMPA. Les données sont celles de NETTALK.*

		SMPA		
		Correct	Faux	Total
JUPA	Correct	95,46	4,54	100,00
		58,73	2,79	61,53
		91,69	7,78	
	Faux	13,83	86,17	100,00
		5,32	33,15	38,47
		8,31	92,22	
Total	64,06	35,95	100,00	
	100,00	100,00		

TAB. 3.9 - *Matrice de confusion entre les transcriptions de JUPA et SMPA. Les données sont celles de NETTALK.*

On note en effet que l'utilisation d'un critère similaire pour ordonner les solutions limite fortement les configurations pour lesquelles la plus simpliste des stratégies de recollement surpasse l'autre, puisque seulement 2,79 % des mots sont bien transcrits par JUPA et incorrectement par SMPA. Compte-tenu de la sophistication introduite dans SMPA, l'existence même de configurations pour lesquelles JUPA se révèle mieux adapté est paradoxale. Ces configurations se justifient toutefois dans le cas particulier de l'anglais, par l'existence d'un nombre important de formes composées dont la prononciation peut se dériver par simple juxtaposition (sans recouvrement) de deux exemples (c.f. la discussion de la section 3.2.5). SMPA, par principe, n'élabore pas de telles constructions (il impose un recollement minimum de taille 1), et ne peut donc tirer parti de ce type particulier de composition. Cette propriété particulière du lexique anglais entraîne un second paradoxe: la supériorité, en termes de justesse, de la plus simple des deux approches analogiques. Pour tous les autres lexiques, toutefois, la supériorité de SMPA est manifeste, aussi bien en ce qui concerne la justesse que la correction: ainsi, les différences moyennes en termes de justesse par mots sont respectivement de plus de 2 points (pour BDLEX et KUN) et de 4 points (pour ILC-20).

nul doute de diminuer sensiblement le taux de silence observé sur ce lexique.

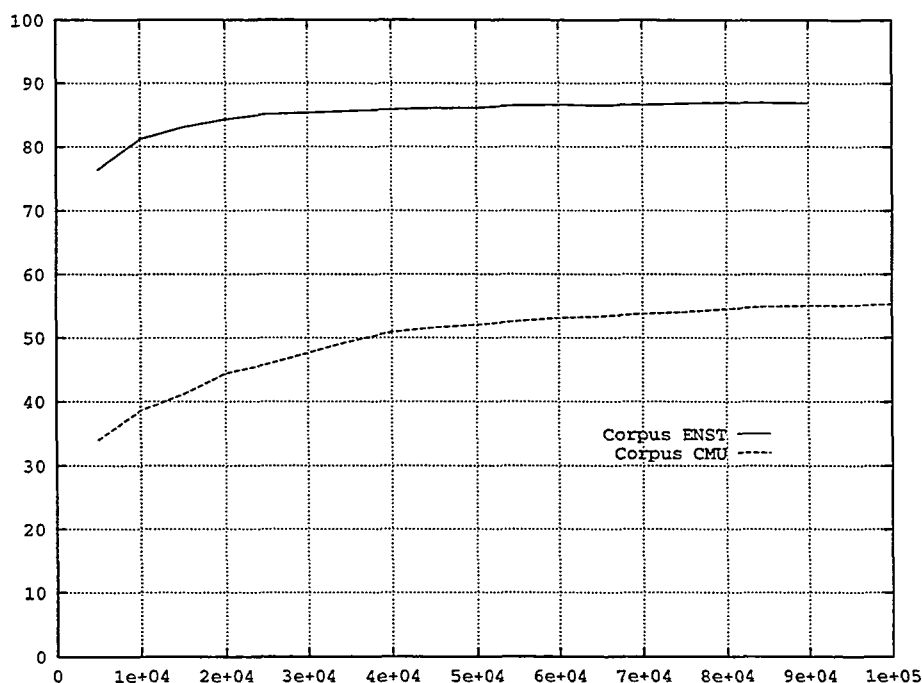


FIG. 3.8 - Courbes d'apprentissage sur les lexiques CMU et ENST

Au total, sur le lexique NETTALK, SMPA améliore les performances en termes de correction par mot de près de trois points, amélioration due, pour une large part, à la quasi-disparition du silence.

Vitesse d'apprentissage et comportement à la limite

La supériorité de SMPA sur les autres stratégies que nous avons évaluées dans les sections précédentes (y compris DEC et PRONOUNCE, dont les évaluations complètes figurent en annexe de ce document) transparait également dans l'allure des courbes d'apprentissage calculées sur les « gros » lexiques dont nous disposons. Ces courbes sont reproduites à la figure 3.8.

Ces courbes sont relativement « plates », et reflètent la capacité de cet algorithme à extraire des généralisations pertinentes dès lors qu'un nombre relativement faible (10 000) d'exemples a été enregistré.

Ceci contribue à rapprocher SMPA des systèmes classificatoires, qui sont capables d'extraire très vite des généralisations, mais pour lesquels la productivité marginale des nouveaux exemples décroît également rapidement. Cette remarque mérite toutefois d'être nuancée : la courbe d'apprentissage obtenue sur le lexique CMU révèle une progression régulière, quoique lente, des performances.

On notera finalement que lorsque l'on utilise la totalité des exemples disponibles, SMPA surpasse DEC sur les deux lexiques, la différence en termes de mots corrects excédant 7 points sur le corpus anglais. Pour donner une seconde

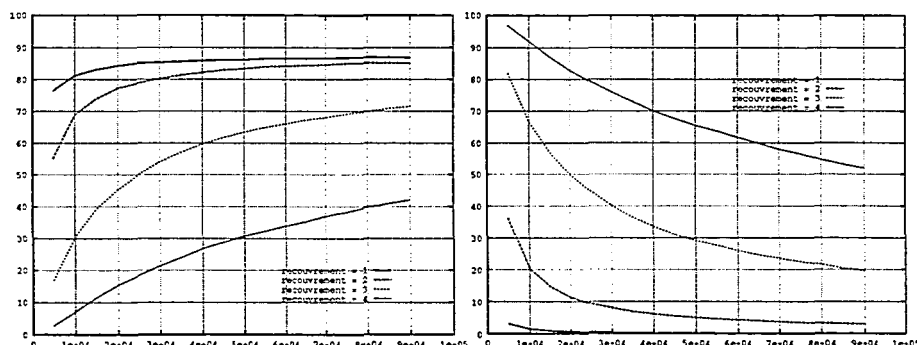


FIG. 3.9 - Évolution des taux de correction et de silence en fonction de la contrainte de recouvrement.

Les données sont celles du lexique ENST.

référence, plus sujette à caution dans la mesure où ces résultats n'ont pas été produits par nos programmes, le pourcentage de mots corrects sur le lexique CMU obtenus par une modélisation markovienne (voir pour une description sommaire la section 2.2.3) sont de 42,8 % (Besling 1995), soit plus de 10 points inférieurs aux nôtres.

Arbitrage silence/correction

Le critère structurel de recouvrement maximal permet enfin de moduler, de manière plus fine que ne le ferait un critère fréquentiel (c.f. la section 2.4.1), le rapport entre silence et correction. Un premier élément étayant cette affirmation donné sur la figure 3.2 (page 93), qui témoigne du bénéfice en termes de correction induit par un contrôle plus strict (i.e. un recouvrement plus grand) des transcriptions. Pour donner plus de poids à cette affirmation, nous avons légèrement modifié SMPA, en introduisant un paramètre supplémentaire, qui correspond à la taille minimale du recouvrement pour qu'une arête soit créée (voir section 3.3.2). Dans la version originale, ce paramètre vaut 1, c'est-à-dire qu'il y a compatibilité et possibilité de recollement entre deux segments dès qu'ils ont un caractère en commun. Le graphique 3.9 reproduit la déformation des courbes de correction (à gauche sur le graphe) et de silence (à droite) lorsqu'on fait varier ce paramètre entre 1 et 4.

Ce graphique met clairement en évidence la manière dont le durcissement de la contrainte de compatibilité permet de jouer sur les taux de correction (et de justesse) et de silence. Chaque augmentation de la contrainte conduit à un « aplatissement » simultané des deux courbes, correspondant à une progression moins rapide des performances, mais à une valeur asymptotique (correspondant à la justesse) légèrement meilleure : en moyenne la justesse est en effet de l'ordre de 87 % (pour $n = 1$), de 88 % (pour $n = 2$), de 90 % (pour $n = 3$), et de 89 % (pour $n = 4$).

3.3.4 Conclusion et perspectives

Nous avons présenté dans cette section un algorithme de prononciation fonctionnant sur le principe de recollement de segments de prononciations attestées, algorithme qui généralise les approches de Byrd et Chodorow d'une part, et de Dedina et Nusbaum d'autre part. En imposant un critère structurel de recouvrement maximal, permettant d'ordonner les différents recollements valides, nous avons montré qu'une telle approche était susceptible de fournir des performances supérieures, sur la tâche de transcription, à ceux obtenus par l'algorithme de Dedina et Nusbaum, et à ceux obtenus en utilisant un algorithme classificatoire.

Cette approche prometteuse pêche néanmoins par deux aspects importants : des erreurs encore trop fréquentes, et une modélisation de l'analogie insatisfaisante. Nous discutons ces deux points dans les paragraphes qui suivent.

Quelques voies d'amélioration

Les résultats auxquels nous sommes parvenus restent néanmoins bien en-deçà niveaux de correction nécessaires à une utilisation en production d'un tel mécanisme pour la transcription. La naïveté du formalisme de représentation utilisé, ainsi que la simplicité des mécanismes de segmentation et de recollement mis en œuvre nous incitent à considérer qu'aux prix de quelques raffinements, des performances nettement supérieures pourraient être atteintes. Suggérons quelques-unes de ces voies d'amélioration :

- la procédure de constitution du dictionnaire de segments fonctionne pour l'instant de manière excessivement brutale, puisque *tous* les segments possibles sont conservés. Il nous semble que des stratégies de nettoyage de ce dictionnaire, fondées soit sur un alignement plus réaliste des chaînes graphiques et phonétiques, soit sur des observations fréquentielles permettraient, sans trop augmenter le silence, de se débarrasser des segments les moins fiables (c.f. la discussion de la section 3.1.2);
- le mécanisme de recollement envisage pour l'instant tous les recollements possibles. Des heuristiques simples peuvent être mises en œuvre pour « sécuriser » les recollements, de manière semblable à ce que nous avons suggéré pour JUPA (voir la section 3.2.4). De manière plus générale, il nous semble possible et désirable d'imposer des contraintes plus spécifiques sur la nature des recollements licites, qui intégreraient des connaissances symboliques du domaine : favoriser les recollements consonantiques par rapport aux recollements vocaliques est un exemple, mais on pourrait également envisager l'intégration de connaissances morphotactiques (autoriser les recollements sans recouvrement en anglais) ou morpho-syntaxiques (la transcription d'un verbe n'utilise que des fragments de verbes, etc);
- la représentation très pauvre des chaînes graphiques et phonémiques gagnerait sans aucun doute à être enrichie. Le principe du recollement est en effet suffisamment souple pour autoriser la prise en compte, en entrée comme en sortie, d'informations additionnelles (frontières de syllabes, de

morphèmes, représentation plus riche des phonèmes), qui pourraient améliorer les performances. Illustrons ce point par deux exemples :

- la représentation des frontières de morphèmes est un autre moyen de « récupérer » des schémas de simple juxtaposition. Si en effet, les lexies *sun* et *light* sont connues, et représentées *+sun+* et *+light+*, alors *sun+light* est recombinaison avec un recouvrement de taille 1 (le caractère marquant la frontière morphématique);
- l'utilisation d'une représentation distribuée des phonèmes sous la forme de vecteurs de traits booléens est également envisageable, et permettrait d'augmenter les capacités de généralisation du mécanisme.

SMPA et l'analogie

Du point de vue de la modélisation de l'analogie, il apparaît que SMPA s'éloigne quelque peu, notamment si on le compare à JUPA, de la ligne générale de réflexion que nous nous sommes fixés. La disparition du silence, en particulier, nous semble être un symptôme de cette divergence : certains mots sont lexicalement plus isolés que d'autres, ne disposent pas d'analogie, et pourtant SMPA parvient à les prononcer tous. Il n'est pas surprenant dans ce contexte que, compte-tenu également de la multiplicité des lexies qui peuvent *in fine* contribuer à la prononciation d'une cible, le concept d'analogie perde de son intelligibilité.

Cette analyse se trouve renforcée si l'on essaie d'exprimer l'approche par recollement de segments dans les termes d'un rapport de proportion. Ainsi reformulée, cette approche ne se distingue en réalité plus tellement des approches classificatoires du type "apprentissage à base d'instances" (c.f. la section 2.4.6). Le principe du recollement se glose en effet par : « le segment *s* est au mot *x* ce que le segment *s* est au mot *y* », illustrant l'incomplétude du mécanisme analogique mis en œuvre, puisque le quatrième terme de l'analogie ne transparaît que par un principe d'identité d'un segment à un autre. Nous aurons l'occasion de revenir plus en détail sur ce point à la section 3.5.

Ces deux remarques nous incitent fortement, en dépit des résultats encourageants que nous avons pu obtenir, et des principes généraux intéressants que les techniques de recombinaison mettent en œuvre, à envisager des modélisations relativement différentes de l'analogie pour la transcription, modélisations auxquelles le chapitre 4 sera consacré.

3.4 Un formalisme probabiliste pour la transcription graphème-phonème par recollement de segments : le modèle multigramme à deux niveaux

Nous abordons dans cette section le problème de la transcription par recollement de segments sous un angle radicalement différent des approches « analogiques » développées dans les deux sections précédentes. Les oppositions sont en effet apparemment tranchées entre le modèle probabiliste des multigrammes à deux niveaux et un algorithme tel que SMPA :

- l’approche analogique tirait parti d’un alignement entre les chaînes graphiques et phonémiques, alors que l’approche probabiliste fonctionne sur des données non-alignées;
- l’approche analogique cherchait à produire une transcription par recombinaison de segments compatibles (i.e. partageant une partie commune), alors que l’approche probabiliste se fonde sur une hypothèse d’indépendance entre segments;
- l’approche analogique ordonnait les solutions sur les bases d’un critère structurel (le recouvrement maximal), alors que la modélisation probabiliste privilégie un classement fondé *in fine* sur un décompte de fréquences.

Au-delà de ces oppositions profondes, ces approches partagent une même vue du mécanisme de transcription, celle de la recombinaison d’unités de taille variable, par opposition à la classification d’unités de taille fixe. Ce faisant, elles rencontrent des problèmes similaires, par exemple le choix d’un dictionnaire de segments optimal. Ce problème, que nous avons évoqué sans vraiment le traiter dans la section précédente, reçoit ici un traitement plus adéquat. Cela nous permettra en retour d’inférer qu’une approche telle que SMPA pourrait en fait fonctionner sans aucun alignement. En ce sens, ces approches se complètent, puisqu’elles apportent, chacune à leur manière, des éléments de solution aux problèmes généraux que posent les approches par segmentation/recombinaison. Nous présenterons, à la fin de ce chapitre, un cadre formel unifié permettant d’exprimer de manière plus claire la forte parenté qui unit ces deux approches, et qui justifiera, *a posteriori*, leur inclusion dans une même famille de techniques.

Cette présentation du modèle multigramme à deux niveaux s’organise comme suit : nous commencerons par présenter le modèle probabiliste qui soutient cette approche, en en marquant le lien de filiation avec le modèle multigramme simple; nous présentons ensuite une batterie de résultats préliminaires, avant de tirer les leçons de cette incursion (qui restera la seule) dans le paradigme de l’apprentissage probabiliste.

3.4.1 Présentation (succincte) du modèle

Le modèle multigramme à deux niveaux⁶ est une extension du modèle multigramme (Bimbot *et al.* 1994; Deligne & Bimbot 1995; Bimbot, Deligne, & Yvon 1995) visant à capturer, dans un formalisme probabiliste, les régularités statistiques identifiables dans l'observation de couples de séquences appariées. Nous présentons ici succinctement le modèle mis en œuvre dans cette approche, en insistant principalement sur les points de recoupements avec les approches exposées dans les sections précédentes. Pour une présentation plus détaillée du modèle, incluant notamment une discussion des questions d'estimation des paramètres, ainsi que l'exposé d'applications de ces modèles au problème du décodage acoustico-phonétique, nous renvoyons aux travaux de Sabine Deligne (1995; 1996).

Retour sur le modèle multigramme

Le modèle multigramme envisage la modélisation des processus de génération de suites de symboles sous la forme de l'émission successive et *indépendante* de segments de taille *variable*. Formellement, cette approche modélise l'émission d'une séquence de symboles comme un processus markovien (voir la section 2.2.3) dont les états (cachés) correspondent aux indices dans la chaîne des points de rupture. À ce titre, ce modèle se distingue singulièrement des modèles *n*-grammes qui modélisent les mêmes observables (des successions de symboles) en postulant que ces chaînes résultent de l'émission successive de symboles atomiques, chaque émission dépendant des émissions précédentes à un horizon de taille fixe (*n*).

Ce modèle vise explicitement à capturer les redondances (acoustiques, phonémiques, graphémiques) identifiables dans des corpus langagiers, et dont il est assumé qu'ils sont la trace de phénomènes essentiellement distributionnels contraints par la phonotactique, la morphologie et la syntaxe. Sous cette hypothèse, un relativement petit nombre de segments (idéalement des morphèmes), dont la taille est variable (dans des limites étroites) doivent alors être identifiables par la simple observation de leur très forte fréquence, à condition qu'on puisse les séparer correctement des unités plus rares (idéalement les racines) avec lesquels ils se recombinent. On trouvera dans (Cartwright & Brent 94) une mise à l'essai alternative de la même hypothèse, exprimée dans les termes de la recherche d'une description de longueur minimale⁷.

La capacité du modèle multigramme à parvenir, *via* la production de segmentations de corpus, à identifier des unités linguistiquement pertinentes a été mise en lumière dans les travaux sus-cités, où une comparaison (avantageuse) avec les modèles *n*-grammes est également effectuée.

6. Le modèle développé ici est le fruit d'une collaboration avec Frédéric Bimbot et Sabine Deligne; d'une rencontre également entre deux problématiques: que cette rencontre ait eu lieu en dit assez sur les convergences potentielles entre les approches mises en œuvre pour traiter le signal de parole et les documents écrits, en dit beaucoup aussi sur la gentillesse et l'ouverture de mes camarades « d'en bas ».

7. Minimal Description Length (MDL).

Le modèle comultigramme

L'extension à deux niveaux de ce modèle se fonde sur une hypothèse équivalente, qui est qu'un corpus de séquences appariées (ici, un texte et sa transcription phonétique) se laisse formellement décrire par un processus d'émission par une source sans mémoire de paires de segments, chaque émission étant indépendante des émissions précédentes.

Formellement, si l'on note $W = (x, y)$ un tel corpus, et L une segmentation conjointe de x et de y , la vraisemblance du corpus dans le modèle s'exprime simplement comme la somme des vraisemblances associées à chacune des segmentations conjointes (nous dirons aussi *cosegmentations*) possibles du corpus :

$$\mathcal{L}(W) = \sum_{L \in \{L\}} \mathcal{L}(W, L) \quad (3.2)$$

Une segmentation conjointe, dans ce modèle, correspond simplement à la donnée d'une série de couples d'indices (i, j) tels que :

$$x = g_1 \dots g_n = (g_1 \dots g_{i_1})(g_{i_1+1} \dots g_{i_2}) \dots (g_{i_p+1} \dots g_n)$$

$$y = p_1 \dots p_m = (p_1 \dots p_{j_1})(p_{j_1+1} \dots p_{j_2}) \dots (p_{j_p+1} \dots p_m)$$

On notera dans la suite γ_i et π_i les i èmes segments dans une cosegmentation L de W . Une telle segmentation étant alors donnée, sa vraisemblance, compte-tenu de l'hypothèse d'indépendance entre les segments successivement émis, s'exprime comme le produit des probabilités d'émission des paires de segments γ_i, π_i , soit :

$$\mathcal{L}(W, L) = \prod_i p \left(\begin{array}{c} \gamma_i \\ \pi_i \end{array} \right) \quad (3.3)$$

L'utilisation de ce modèle en génération consiste à inférer, les probabilités d'émission étant connues, la séquence d'émissions (formellement équivalente à une segmentation conjointe) qui a le plus vraisemblablement possible produit le corpus observé, soit L^* définie par :

$$L^* = \operatorname{argmax}_{L \in \{L\}} \mathcal{L}(W, L)$$

Pour rendre ce modèle d'une part identifiable, et d'autre part capable de rendre compte d'un certain nombre de faits linguistiques, il faut naturellement le contraindre en imposant des tailles maximales aux segments émis. Sans cela, le nombre de paramètres devient infini, d'où des problèmes certains d'identifiabilité; et le modèle inintéressant : la meilleure segmentation conjointe est tout simplement celle qui reconnaît le corpus comme un tout monolithique. On notera dans la suite du texte μ (respectivement ν) la taille maximale pour un segment de x (resp. de y), et l'on désignera sous le terme de modèle (μ, ν) un modèle respectant ces contraintes de taille.

L'exemple suivant, tiré de (Deligne, Yvon, & Bimbot 1995) permet de visualiser dans un cas très simple les termes de cette expression : supposons que notre

Prononcer des mots inconnus

Décodage simple Utiliser ce modèle pour la transcription graphème-phonème⁸ suppose la construction, étant donnée simplement la chaîne x , d'une règle de décision qui permette de décoder la chaîne y manquante, correspondant censément à $\phi(x)$. Dans le cadre formel que nous avons choisi, la manière la plus naturelle consiste à choisir y de façon à maximiser la vraisemblance conditionnelle de y sachant x , soit :

$$y^* = \operatorname{argmax}_{\{y\}} \mathcal{L}(y | x)$$

En décomposant cette vraisemblance comme la somme sur toutes les segmentations possibles, puis en faisant l'hypothèse que seule la segmentation optimale L^* contribue significativement à cette somme, il vient :

$$y^* = \operatorname{argmax}_{\{y\}} \sum_{\{L\}} \mathcal{L}(y, L | x) \quad (3.4)$$

$$= \operatorname{argmax}_{\{y\}} \mathcal{L}(y, L^* | x) \quad (3.5)$$

$$= \operatorname{argmax}_{\{y\}} \mathcal{L}(y, L^*, x) \quad (3.6)$$

Décodage avec modèle Une manière alternative de procéder consiste à remarquer que la chaîne ϕ qui réalise le maximum de $\mathcal{L}(y | x)$ réalise également, x étant fixé, le maximum de la vraisemblance conjointe. En décomposant alors cette vraisemblance conjointe par la règle de Bayes en $\mathcal{L}(x | y)\mathcal{L}(y)$, on peut, moyennant la même approximation que précédemment sur la vraisemblance de la segmentation optimale, dériver :

$$y^* = \operatorname{argmax}_{\{y\}} \sum_{\{L\}} \mathcal{L}(x, L | y)\mathcal{L}(L, y) \quad (3.7)$$

$$= \operatorname{argmax}_{\{y\}} \mathcal{L}(x, L^* | y)\mathcal{L}(L^*, y) \quad (3.8)$$

Cette décomposition-ci a le mérite de faire apparaître un second terme : en sus de $(x, L^* | y)$, qui correspond à la qualité de l'appariement entre y^* et x , un second terme entre en jeu, à savoir $\mathcal{L}(L^*, y)$, qui prend en compte de la vraisemblance de y^* en tant que séquence de $\mathcal{P}$, et ce indépendamment de toute contrainte d'appariement. Pour estimer ce dernier terme, nous avons utilisé un modèle bigramme portant sur les séquences de segments phonétiques. Nous référerons en conséquence, dans la suite du texte, à cette technique de décodage de la séquence x sous le nom de *décodage avec modèle*.

8. Une application plus immédiate de ce modèle est de l'utiliser pour en extraire des alignements plurivoques des deux séquences, alignement dont les algorithmes décrits dans les sections initiales ce chapitre pourraient directement tirer parti. Les quelques expériences préliminaires menées dans cette direction-là se sont toutefois révélées insuffisamment probantes pour pouvoir valider en l'état les capacités de ce modèle à accomplir cette tâche.

Lien avec SMPA Comme dans SMPA, la procédure ici mise en œuvre, consiste, du moins dans le cas du décodage sans modèle, à répertorier les appariements observés sur des segments d'éléments du lexique d'apprentissage; à affecter un score à chacun de ces cosegments; puis à extraire de cet ensemble la « meilleure » séquence de cosegments compte-tenu de la cible à transcrire. Les points de divergence entre ces deux approches portent principalement sur les deux aspects suivants :

- *Contenu du dictionnaire* : Dans le cas de SMPA, les scores associés aux segments correspondent simplement à la fréquence d'observation de l'appariement. En particulier, si un segment graphique est associé à plusieurs segments phonétiques différents, cette ambiguïté est enregistrée dans le dictionnaire. Ce n'est pas toujours le cas dans le modèle de multigrammes à deux niveaux, puisque dans le cas du décodage simple, seul l'appariement le plus probable est conservé. En effet, compte-tenu de l'équation 3.6, si les paires de segments $\begin{pmatrix} \gamma_i \\ \pi_j \end{pmatrix}$ et $\begin{pmatrix} \gamma_i \\ \pi'_j \end{pmatrix}$ sont toutes les deux observées, la première avec une probabilité supérieure à la seconde, alors n'importe quelle segmentation d'une chaîne contenant γ_i appariée avec π_j sera plus probable qu'une autre appariant γ_i avec π'_j . Dans la mesure où le modèle ne considère que la segmentation la plus probable, conserver la paire $\begin{pmatrix} \gamma_i \\ \pi'_j \end{pmatrix}$ est inutile.

À l'inverse, le dictionnaire de cosegments utilisé par le modèle probabiliste est plus complet que son homologue en ce qu'il contient virtuellement tous les cosegments possibles dont le composant graphémique n'est pas déjà enregistré : comme il est fréquent dans ce type de modèle, ces cosegments virtuels sont affectés d'une probabilité très faible, mais non-nulle. SMPA, lui, est incapable de restituer des appariements non-attestés sur le lexique d'apprentissage.

L'incidence sur le contenu des dictionnaires induite par le fait que SMPA utilise des lexiques alignés, alors que le modèle multigramme fonctionne de manière totalement non-supervisée nous semble par contre faible : l'alignement n'est pour SMPA qu'un moyen simple de se débarrasser *a priori* de paires de segments non-pertinents, paires qui disparaissent sur des bases statistiques lors de l'apprentissage du modèle.

- *Sélection du meilleur recollement* : La seconde différence majeure entre ces deux approches porte sur la manière de sélectionner la meilleure prononciation. Dans l'approche multigramme, cette sélection résulte d'un compromis dynamique entre les différentes segmentations possibles de x , chaque partie de la segmentation finalement élue fournissant le fragment phonémique qui lui correspond le plus probablement, indépendamment⁹ des autres composants de la segmentation, *indépendamment en particulier des éléments qui lui sont adjacents*. Dans SMPA au contraire, c'est sur la base d'une dépendance maximale avec la production associée au segment précédent que chaque segment graphique participant à la solution est choisie : la

9. Du moins dans la stratégie de décodage sans modèle.

transcription résulte non-pas d'une segmentation, mais d'un recouvrement par des segments d'intersection maximale.

3.4.2 Évaluation sur le lexique bdlex

Dans cette section, nous présentons les résultats d'expériences préliminaires conduites avec le seul lexique BDLEX. Un des principaux problèmes posé par l'estimation de modèles probabilistes par les procédures d'estimations que nous avons employées est celui de leur convergence vers un optimum local. La manière dont est conduite la phase d'initialisation des paramètres du modèle est, dans ce contexte, cruciale, puisqu'elle conditionne fortement la valeur finale des paramètres qui va être atteinte.

Avec d'une part le problème de la recherche de la meilleure stratégie d'apprentissage; d'autre part la détermination des meilleures contraintes de tailles, et enfin l'évaluation de deux stratégies de décodage (avec ou sans modèle auxiliaire), nous disposons dans les faits d'un grand nombre de degrés de liberté pour conduire nos évaluations. Dans ces conditions, la recherche par tâtonnement de la meilleure stratégie globale (contraintes, initialisation, décodage) est longue et fastidieuse. C'est de cette recherche dont nous avons souhaité rendre compte dans le cas particulier du lexique BDLEX, en espérant que les variations des performances du modèle induites par les différentes stratégies mises en œuvre mettent mieux en relief les forces et faiblesses de cette approche que ne l'aurait fait une évaluation multilingue moins complète. Pour ce motif, le protocole expérimental diffère sensiblement de celui, plus général, décrit dans l'annexe D. Nous commencerons par décrire ce protocole, avant de donner quelques mesures de l'influence des paramètres de taille μ et ν , des diverses stratégies d'écritage du dictionnaire, et des diverses techniques de décodage.

Protocole Expérimental

Tous les résultats qui vont suivre ont été obtenus par application du protocole expérimental suivant, qui comprend trois étapes principales :

- *Initialisation* : Pendant cette étape, le dictionnaire de tous les cosegments possibles compte-tenu des contraintes de taille est constitué, et les probabilités initiales, calculées. Comme on peut le voir sur l'exemple précédent du lexique (*trois*, /tʁwa/), ce recensement exhaustif conduit à l'enregistrement d'un très grand nombre de cosegments, qui, dans une large mesure, correspondent à des appariements fort hasardeux (ainsi $\begin{pmatrix} tro \\ t \end{pmatrix}$). Dans la mesure où l'on peut penser que ces appariement-là d'une part ralentissent, et d'autre part brisent l'apprentissage, il est important de s'en débarrasser le plus tôt possible.

Dans cette optique, deux démarches ont été explorées : l'une fait l'hypothèse que les cosegments les plus hasardeux sont les cosegments les plus rares, et procède par élimination de ceux dont la fréquence est au-dessous d'un certain seuil. Une autre façon de faire est de se donner une cosegmentation initiale, par exemple résultant d'une application antérieure du

modèle, et de n'enregistrer que les cosegments qui permettent de ne pas trop s'éloigner de la segmentation de départ.

- *Apprentissage* : Une fois le dictionnaire créé, on effectue durant cette étape 10 itérations de l'algorithme d'apprentissage, chaque itération consistant en une étape de segmentation optimale compte-tenu du dictionnaire et des probabilités de départ, suivie d'une étape de reconstruction du dictionnaire à partir de la segmentation obtenue. À la fin de chaque itération, un mécanisme supplémentaire de réduction du dictionnaire sur des bases fréquentielles est optionnellement mis en œuvre. Le nombre d'itération a été fixé *a priori* à 10 sur la base d'observations empiriques de la vitesse de convergence.
- *Transcription* : Sur la base du recensement ainsi effectué, un ensemble d'environ 2000 mots non-utilisés pour l'apprentissage est transcrit, en utilisant l'une ou l'autre des stratégies de décodage définies en 3.4.1. Ces transcriptions sont comparées aux transcriptions de référence, pour fournir des scores de transcription en termes de pourcentage de mots et de phonèmes corrects, d'une manière tout à fait similaire aux procédures utilisées dans nos autres évaluations.

Précisons que dans les paragraphes qui suivent, sauf mention explicite du contraire, les résultats qui sont reportés ont été obtenus en utilisant un principe de décodage sans modèle auxiliaire.

Influence des contraintes de taille

Dans ce modèle de segmentation conjointe, les deux paramètres les plus importants sont ceux qui concernent la taille maximale autorisée des segments dans l'une ou l'autre des chaînes, puisque ces paramètres conditionnent les dimensions du modèle à estimer. Il est à noter que pour des valeurs arbitraires de ces paramètres, il n'est pas garanti qu'un corpus $W = (x, y)$ puisse être segmenté de manière à respecter ces contraintes. L'exemple le plus immédiat d'une telle impossibilité est celui où les deux paramètres tout deux sont pris égaux à 1 : tout corpus pour lequel $l(x) \neq l(y)$ ne peut en effet être segmenté par appariement terme à terme de segments de longueur 1.

La taille 3.11 présente un tableau complet des performances en décodage du modèle lorsque μ varie entre 2 et 6, et ν varie entre 1 et 6. Dans ces expériences, aucune procédure de réduction de la taille du dictionnaire n'est mise en œuvre.

La tendance qui se dégage de ces premiers résultats est que les performances s'améliorent au fur et à mesure que l'on augmente la taille des segments licites, du moins tant que les deux valeurs restent proches. On notera que deux points n'ont pu être calculés à cause de l'explosion combinatoire du nombre de cosegmentations possibles. Une autre vue, plus synthétique, des mêmes données est reproduite sur la figure 3.10, où l'on a représenté les performances en termes de mots corrects en fonction des deux paramètres de taille.

Modèle	Correction		Modèle	Correction	
	% mots	% symboles		% mots	% symboles
2,1	25,59	76,51	4,3	54,46	89,48
2,2	34,01	80,80	4,4	44,92	87,64
2,3	34,59	81,10	4,5	54,72	89,40
2,4	35,12	81,24	4,6	53,44	88,52
2,5	35,12	81,28	5,1	26,65	77,36
2,6	35,17	81,26	5,2	47,18	86,53
3,1	27,05	77,59	5,3	57,43	89,93
3,2	43,95	85,91	5,4	44,39	87,55
3,3	47,63	87,51	5,5	56,14	90,14
3,4	37,78	84,59	5,6	49,98	87,76
3,5	47,10	87,02	6,1	26,65	77,36
3,6	47,41	87,02	6,2	39,02	83,67
4,1	26,65	77,36	6,3	57,29	89,57
4,2	47,01	86,57	6,4	53,66	89,00

TAB. 3.11 – Performances en décodage pour différentes valeurs des contraintes de taille

% mots corrects

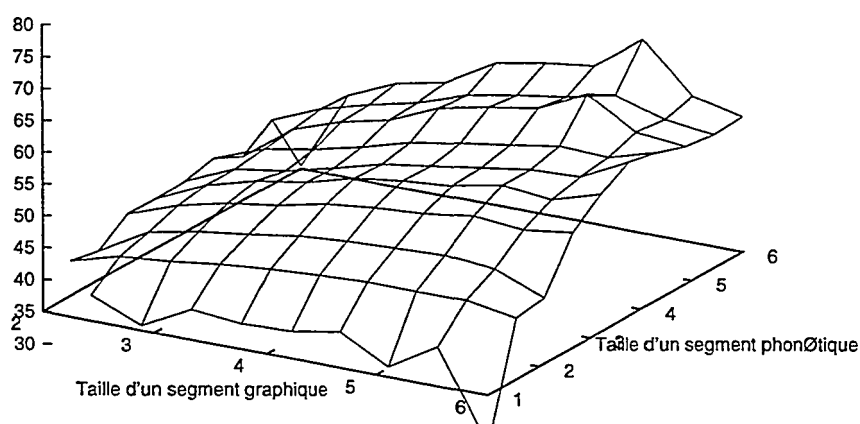


FIG. 3.10 – Performances sur la tâche de transcription en fonction des contraintes de taille

Prononciation par analogie : motivations, formalisations, évaluations

Écrêtage	Correction		Écrêtage	Correction	
	% mots	% symboles		% mots	% symboles
0-0	57,43	89,93	3,5	56,23	89,48
0-1	59,33	90,39	3,10	50,86	87,70
0-3	56,10	89,55	5,0	59,25	90,61
0-5	55,25	89,19	5,1	59,51	90,72
0-10	49,71	87,16	5,3	58,67	90,34
1-0	58,76	90,40	5,5	58,00	90,10
1-1	60,44	90,83	5,10	53,53	88,41
1-3	57,34	89,92	10,0	58,45	90,15
1-5	55,96	89,55	10,1	58,67	90,23
1-10	50,02	87,27	10,3	58,98	90,24
3-0	59,07	90,60	10,5	56,98	89,73
3-1	59,91	90,72	10,10	55,57	89,33
3-3	58,36	90,38			

TAB. 3.12 – Performances en décodage pour le modèle (5,3) pour différents seuils d'écrêtage

Diverses stratégies d'apprentissage

Dans cette section, nous nous intéressons plus particulièrement aux différentes méthodes susceptibles d'améliorer la qualité de l'initialisation et la vitesse de l'apprentissage.

Écrêtage du dictionnaire par fixation de seuils Pour chacun de ces modèles, nous avons tout d'abord évalué les performances de diverses stratégies de réduction du dictionnaire des co-segments. Ces stratégies sont définies par le choix de deux nouveaux paramètres : c , qui correspond au seuil de fréquence minimale pour qu'un co-segment soit conservé dans le dictionnaire à l'issue de l'étape d'initialisation, et d , qui joue un rôle similaire durant les itérations de la procédure d'estimation. Les performances en décodage d'un modèle (5,3) pour différentes valeurs de ces deux seuils sont reportées dans la table 3.12.

La sensibilité des résultats à ces différents seuils est relativement faible, puisque tous les scores sont compris dans une fourchette de l'ordre de 2 points. Il semble toutefois que dans tous les cas, il soit toujours préférable, à c fixé, de choisir d plus faible; et à d fixé, de choisir c plus élevé. Des résultats confirmant cette règle empirique ont été obtenus pour d'autres valeurs des contraintes de taille (modèles (2,2), (3,3), (4,3), ...).

Initialisation à partir d'une cosegmentation L'idée est ici d'estimer un premier modèle relativement contraint, à l'aide duquel on pourra produire une première cosegmentation approximative du lexique d'exemples. Nous avons pour cela eu recours à un modèle (2,2). Au cours de l'étape d'initialisation du dictionnaire des cosegments, on utilise ces appariements initiaux pour restreindre l'ensemble des cosegmentations considérées. Pour ce faire, on introduit une contrainte qui exprime que les appariements à prendre en compte préservent

Modèle	Modèle (2,2)		Phonogrammes	
	% mots	% symboles	% mots	% symboles
3,3	48,87	87,72	50,11	88,04
4,3	59,33	90,59	60,13	90,80
5,3	60,93	90,85	62,04	91,06
5,5	61,73	91,52	62,79	91,73
6,5	59,60	90,66	60,93	91,03
6,6	58,40	90,39	59,91	90,68
7,5	56,41	89,74	58,40	90,26
7,6	52,99	88,84	53,92	89,06
7,7	51,35	88,39	52,20	88,54

TAB. 3.13 – Performances en décodage pour différentes contraintes de taille, avec initialisation.

approximativement les frontières de cosegments. Si on dispose originellement d'un alignement $\begin{pmatrix} ab \\ \alpha \end{pmatrix}$, alors les seuls cosegments licites commençant par un a devront impérativement commencer par un $/\alpha/$; et de même pour ceux qui commencent par un b . Dans des cas plus complexes, tels que $\begin{pmatrix} abcd \\ \alpha\beta \end{pmatrix}$, un principe de proportionnalité s'adjoint pour contraindre les « nouveaux » cosegments commençant par a ou b à débiter par un $/\alpha/$; et ceux commençants par c ou d à s'aligner sur le $/\beta/$.

Supposons par exemple que l'alignement initial du mot *trois*, construit en utilisant un modèle (2,2) soit $\begin{pmatrix} tr \\ t\mathfrak{r} \end{pmatrix} \begin{pmatrix} o \\ w \end{pmatrix} \begin{pmatrix} is \\ a \end{pmatrix}$, alors les seules segmentations additionnelles qui seront considérées lors de l'initialisation d'un modèle (3,3) sont les suivantes : $\begin{pmatrix} tro \\ t\mathfrak{r}w \end{pmatrix} \begin{pmatrix} is \\ a \end{pmatrix}$, $\begin{pmatrix} t \\ t \end{pmatrix} \begin{pmatrix} roi \\ \mathfrak{r}w \end{pmatrix} \begin{pmatrix} s \\ a \end{pmatrix}$ et $\begin{pmatrix} tr \\ t\mathfrak{r} \end{pmatrix} \begin{pmatrix} ois \\ wa \end{pmatrix}$.

Dans l'ensemble, cette technique a permis d'améliorer de manière significative les performances obtenues avec une initialisation exhaustive, comme en témoignent les résultats listés dans la table 3.13.

L'autre avantage de cette stratégie de réduction de la taille du dictionnaire est qu'elle s'effectue *a priori*, et autorise ainsi l'estimation de modèles « plus gros » que précédemment. Nous pouvons ainsi constater, que, comme la table 3.11 nous le laissait supposer, il semble exister une valeur critique des paramètres de taille au-delà de laquelle les performances se dégradent. Dans notre cas, les meilleurs résultats sont obtenus avec un modèle (5,5).

La validité de cette stratégie est finalement confortée par les résultats obtenus en utilisant pour l'initialisation du modèle la cosegmentation en phonogrammes (Catach 1984) qui est explicitement disponible dans les données de BDLEX (voir l'annexe F). Ces résultats, qui figurent dans la troisième colonne du tableau, sous le chapeau « phonogrammes », sont en effet légèrement meilleurs que ceux obtenus en initialisant à partir du modèle (2,2).

Les deux techniques de décodage

L'intégralité des évaluations (en faisant varier les paramètres de taille, en écrétant de moult façons, en initialisant à partir d'une cosegmentation) que l'on vient de présenter a été répétée en utilisant la stratégie alternative de décodage. Sans présenter en détail l'intégralité des résultats, disons simplement que dans l'ensemble, ces expériences complémentaires ont montré que :

- pour cette technique de décodage également, il est avantageux de mettre en œuvre des procédés de réduction de la taille du dictionnaire, y compris pendant les itérations d'apprentissage;
- l'utilisation d'un modèle auxiliaire permet d'améliorer marginalement le meilleur des résultats obtenu : 93,72 % de phonèmes corrects avec un modèle (3,2) et modèle auxiliaire contre 92,31 % avec un modèle (5,5) sans modèle auxiliaire. Plus généralement, ce raffinement de la technique de décodage permet de doper fortement les performances des « petits » modèles, puisque pour un modèle (3,3) par exemple, les scores en termes de mot corrects s'améliorent de près de 4 points (91,24 % contre 87,51 % de phonèmes corrects) quand on adjoint un modèle de séquences.

3.4.3 Discussion et Perspectives

En dépit de nos efforts, les performances du modèle de multigrammes à deux niveaux, bien qu'honorables, sont pour l'instant bien en-deçà de celles des algorithmes que nous avons présenté au cours de ce chapitre, comme de celles de DEC. Une des premières raisons que l'on peut invoquer pour expliquer ces différences est que DEC, comme la totalité des algorithmes présentés dans les sections précédentes, apprend à partir de données alignées, ce qui n'est pas le cas du modèle comultigramme. L'étude de transcriptions typiques (voir infra), qui ne révèle qu'un nombre très marginal d'erreur d'alignement, nous incite toutefois à approfondir cette analyse. Parmi les diverses hypothèses alternatives considérées pour expliquer alors cette faiblesse relative des résultats, deux semblent particulièrement plus particulièrement dignes d'être discutées, qui font l'objet des paragraphes qui suivent. Nous concluons ensuite cette section par une note plus optimiste sur la pertinence linguistique des segmentations produites.

Une hypothèse que par contre nous avons rejeté est celle de la piètre qualité de l'initialisation qui conduirait à des modèles mal estimés : même lorsque l'on part d'une cosegmentation très précise comme celle que fournit la décomposition en phonogrammes, les résultats restent insatisfaisants. Compte-tenu par ailleurs de l'exploration relativement extensive des différents degrés de libertés dont nous disposons pour sélectionner les valeurs initiales des paramètres (y compris en expérimentant des stratégies hybrides, qui font appel conjointement à une cosegmentation initiale et à des seuils d'écrtage), nous sommes relativement confiant dans notre affirmation qu'une meilleure initialisation (par exemple en utilisant une procédure incrémentale à plus de deux étages) n'améliorerait que marginalement les résultats que nous avons présentés.

Mot	Transcription	Graphie Segmentée	Transcription segmentée
* canitie	/kanisi/	can-it-ie	/kan-i-i/
* crayeux	/kʁɛjø/	cra-y-eux	/kʁa-i-ø/
* fuchsia	/fyʃja/	fu-ch-s-ia	/fy-ʃ-s-ja/
* nausée	/noze/	na-us-ée	/na-us-e/
* regrossir	/ʁəɡʁosiʁ/	re-gr-os-s-ir	/ʁə-gʁ-oz-s-iʁ/
* ruffian	/ʁyʃjɑ̃/	ru-f-fi-an	/ʁy-f-fi-ɑ̃/
* tarots	/taʁo/	tar-ot-s	/taʁ-o-s/
* tunnel	/tynɛl/	tu-n-nel	/ty-n-nɛl/
* unionisme	/ynjonismə/	uni-on-i-sme	/yni-ɔ̃-i-smə/
bougeoir	/buʒwaʁ/	bou-ge-oir	/bu-ʒ-waʁ/
chansonnette	/ʃɑ̃sɔ̃nɛtə/	chan-s-onn-ette	/ʃɑ̃-s-ɔ̃n-ɛtə/
formule	/fɔʁmylə/	for-m-ule	/fɔʁ-m-ylə/
pendard	/pɑ̃daʁ/	pend-ard	/p-ɑ̃d-aʁ/
ponton	/pɔ̃tɔ̃/	pon-ton	/pɔ̃-tɔ̃/
promener	/pʁɔmənɛ/	pro-me-ner	/pʁɔ-mə-ne/
rection	/ʁɛksjɔ̃/	rec-tion	/ʁɛk-sjɔ̃/
relaxe	/ʁələksə/	re-la-xe	/ʁə-la-ksə/
roulotte	/ʁulɔtə/	rou-l-otte	/ʁu-l-ɔtə/
transposer	/tʁɑ̃spoze/	tran-sp-oser	/tʁɑ̃-sp-oze/
trépasser	/tʁɛpase/	tré-pass-er	/tʁɛ-pass-e/

TAB. 3.14 – Quelques transcriptions produites par application du modèle (4,3).

Quelques transcriptions typiques

Afin d'illustrer notre propos par des exemples concrets, nous avons listé quelques productions typiques des erreurs et des succès de l'approche probabiliste (le modèle utilisé est un modèle (4,3) avec écrêtage 10 – 10) dans les tables 3.14 (décodage simple) et 3.15 (décodage avec modèle auxiliaire; seules les entrées qui diffèrent de la table précédente sont reproduites). Les deux premières colonnes de ces tableaux contiennent la graphie et la prononciation associées à une entrée lexicale à transcrire, et les deux suivantes la segmentation et la transcription fournies par le modèle; une étoile signale les erreurs du modèle.

Indépendance et ambiguïté

Une première critique particulièrement fondée est celle de l'hypothèse de fond du modèle, qui considère les émissions successives de segments appariés comme étant indépendantes. Cela simplifie clairement les procédures d'estimation et de transcription, puisque, nous l'avons signalé plus haut, cela permet de ne conserver dans le dictionnaire, pour une chaîne orthographique donnée que le plus probable des segments phonétiques associés.

L'examen de la table 3.14 révèle toutefois tous les dégâts que cette approche du problème provoque: prononciation inconditionnelle du *it* de *canitie*, du *na* de *nausée*, du *-ion* de *unionisme*, mauvaise interprétation des consonnes doubles de *ruffian* et de *regrossir*...

Mot	Transcription	Graphie Segmentée	Transcription segmentée
* canitie	/kanisi/	ca-ni-t-ie	/ka-ni-t-i/
* fuchsia	/fyʃja/	fu-ch-s-ia	/fy-k-s-ja/
* pependard	/pūdaʁ/	p-end-ard	/p-āḍ -aʁḍ/
* ruffian	/ʁyʃjā/	r-u-ff-ia-n	/ʁy-f-ja-n/
* tarots	/taʁo/	tar-ot-s	/taʁ-o-z/
* trépasser	/tʁɛpase/	tré-p-asse-r	/tʁɛ-p-asə-ʁ/
* unionisme	/ynjonismə/	un-ion-i-sme	/yn-jō-i-smə/
crayeux	/kʁɛjø/	cr-a-y-eux	/kʁ-ɛ-j-ø/
nausée	/noze/	n-au-s-ée	/n-o-s-e/
rection	/ʁɛksjō/	r-ec-tion	/ʁ-ɛk-sjō/
regrossir	/ʁɛgʁosiʁ/	re-g-ro-ss-ir	/ʁɛ-g-ʁo-s-iʁ/
relaxe	/ʁɛlaksə/	rel-ax-e	/ʁɛl-aks-ə/
transposer	/tʁāspoze/	tran-s-p-oser	/tʁā-s-p-oze/

TAB. 3.15 – Quelques transcriptions produites par application du modèle (4,3) avec modèle auxiliaire.

L'utilisation d'un modèle de langue auxiliaire permet de faire disparaître, en introduisant une dépendance entre les séquences de segments phonétiques qui permet la prise en compte effective des ambiguïtés, une partie de ces problèmes. Dans la deuxième liste en effet (voir la table 3.15), certains digraphes (notamment les consonnes doubles et la séquence au) sont correctement identifiés, et *unionisme* est partiellement corrigé (le premier *i* est correctement transcrit /j/).

Paradoxes de cosegmentation

Un certain nombre d'erreurs subsistent toutefois, qui proviennent à notre avis de l'inadéquation de formulation du modèle, qui le conduit à affronter de véritables paradoxes de segmentation, comme l'illustre le mot *mangeur*. Considérons en effet quelques segmentations possibles de cette chaîne graphique :

- (1) *ma-ngeur*
- (2) *man-geur*
- (3) *mang-eur*
- (4) *mange-ur*
- (5) *mangeu-r*

Le problème de la première solution, c'est qu'elle isole le *a*, conduisant potentiellement à une transcription /ma/ du premier segment. Pour ne pas se tromper sur l'interprétation de ce graphème, il faut en fait intégrer jusqu'au *g* dans le premier segment. Dans le même temps, la prononciation du *g* lui-même est problématique, et interdit qu'on le sépare du *e* qui suit. Enfin, introduire une frontière de segment dans la séquence finale *eur* est également proscrit. Si alors les contraintes de taille interdisent la segmentation *m-angeur*, ce mot a de fortes chances d'être incorrectement transcrit.

Augmenter toutefois la taille des cosegmentations licites induit également des effets pervers manifestes : les « longs » cosegments étant dans l'ensemble

plus rares et moins ambigus que les petits, ils sont mécaniquement affectés de probabilités plus fortes; par ailleurs l'effet multiplicatif du calcul de la vraisemblance conduit notre modèle à « préférer » les cosegmentations qui comportent peu de termes, le tout conduisant à une bien trop forte utilisation de cosegments « longs » dans les transcriptions. C'est d'ailleurs le problème majeur que nous avons rencontré dans l'utilisation du modèle multigramme à deux niveaux pour produire des alignements plurivoques: alors que dans la distribution des tailles des cosegments donnés par la représentation en phonogrammes, les appariements (1, 1) comptent pour plus de 80 % des cosegments, un modèle (3, 5) par exemple en reconnaît moins de 10 % (et plus de 30 % d'appariements (3, 3)).

À propos des cosegmentations produites

Ces critiques du modèle multigramme à deux niveaux méritent toutefois d'être nuancées: des segmentations et des transcriptions tout à fait pertinentes et correctes sont également produites, telles que celles listées dans les tables précédentes, ce qui témoigne de la relative capacité du modèle à reconnaître des segments interprétables linguistiquement. En particulier, l'identification correcte de certains phonogrammes (les consonnes doubles dans *ruffian* et *regrossir*); les digraphes *au* (dans *nausée*) et *ch* (dans *fuchsia*) est particulièrement encourageante.

De même, la relative capacité de ce modèle à identifier des frontières de segments qui soient morphologiquement interprétables, comme dans *roulotte* ou *unionisme* est une propriété intéressante. Il semble même que l'identification de tels segments soit plus satisfaisante que dans le modèle multigramme simple. Une hypothèse acceptable est à notre sens qu'un segment fréquent est identifié dans l'approche à deux niveaux par une récurrence des mêmes formes graphiques associées aux mêmes formes phonétiques; cette condition de régularité conjointe permettant de filtrer une bonne part des régularités graphiques accidentelles. L'utilisation conjointe des formes graphiques et phonétiques serait donc une meilleure base d'apprentissage des régularités morphologiques que l'une ou l'autre des formes considérée seule. Ce dernier point reste toutefois à confirmer par des expériences spécifiquement consacrées à l'étude des performances en termes de découpage morphologique.

3.5 Conclusion sur les approches par recollement de segments

Nous avons exploré, dans ce chapitre, la validité d'approches de la transcription fondées sur l'utilisation de larges segments de lexies. Cette exploration nous a permis de présenter trois algorithmes exploitant cette idée, avec d'un côté les approches « analogiques » (JUPA et SMPA), de l'autre une approche purement statistique (le modèle multigramme à deux niveaux). Les premiers enseignements que nous pouvons tirer de ces expériences nous permettent de réaffirmer notre optimisme quant à la pertinence de ce type d'approche vis-à-vis de la transcription. Les performances de SMPA, en particulier, montrent que de

tels algorithmes sont susceptibles de se montrer plus justes que des approches classificatoires. De surcroît, nous avons mis en évidence une caractéristique, importante à nos yeux, des algorithmes « analogiques », à savoir leur capacité à discriminer, de façon relativement pertinente, certaines chaînes pour lesquelles, en l'absence d'indication suffisante, aucune transcription n'est produite. De cette manière, ces algorithmes apportent un élément de réponse à l'une des critiques adressée aux approches classificatoires (voir la section 2.4.2).

Il est de fait possible d'unifier formellement l'ensemble des approches présentées dans ce chapitre, en les réexprimant dans le cadre de la théorie des langages. Cette reformulation s'énonce de la manière suivante. Dans l'ensemble des approches par recollement de segments, la base d'exemple est compressée sous la forme d'un dictionnaire $\mathcal{S}$ de co-segments, c'est à dire de couples (segment graphique, segment phonétique). Dans la mesure où l'on impose des contraintes sur la taille des segments, cet ensemble est fini. Il est alors possible de l'identifier à un alphabet.

Considérons maintenant les alphabets spécifiques, ainsi que les ensembles de chaînes sur cet alphabet reconnus par les différentes techniques que nous avons proposées et/ou évaluées :

- dans le cas du modèle de multigramme, l'alphabet $\mathcal{S}$ correspond à tous les appariements possibles d'une séquence de taille inférieure ou égale à μ dans le domaine graphique avec une séquence de taille inférieure ou égale à ν dans le domaine phonétique. Le langage reconnu sur cet alphabet est tout simplement $\mathcal{S}^*$, puisque toutes les séquences de co-segments sont, dans ce modèle, licites. L'automate correspondant à ce langage est déterministe, mais les ambiguïtés de segmentations font que plusieurs séquences de transitions dans l'automate peuvent correspondre à un mot (non-ségmenté) du lexique. La stochastisation de cet automate, par assignation de probabilités aux transitions, permet de plaquer un ordre (celui du induit par le maximum de vraisemblance) sur les parcours correspondant à une même entrée.
- dans le cas des algorithmes qui imposent un recouvrement entre segments adjacents (SMPA et PRONOUNCE), les alphabets sont restreints aux seuls appariements observés sur le lexique d'exemple compatibles avec l'alignement bi-univoque des lexies avec leur phonétique. Le langage reconnu sur cet alphabet est également beaucoup plus contraint, puisque la condition de recouvrement limite drastiquement les successions de symboles licites. Il est possible de construire explicitement l'automate correspondant à ce langage, en procédant d'une manière très similaire à celle utilisée pour la construction des graphes locaux d'analogie (voir la section 3.3.2) :
 - on associe à chaque symbole du dictionnaire un état, et on marque les états correspondant à des appariements terminaux comme étant des états finaux de l'automate;
 - on ajoute un état initial;
 - on crée une transition entre l'état initial et chaque état correspondant à un appariement initial. Chaque transition ainsi créée correspond à l'émission du symbole (i.e. de la paire de segments) correspondant;

- on crée une transition entre deux états « internes » si et seulement si les cosegments auxquels ils correspondent sont compatibles, au sens où nous avons défini la compatibilité à la section 3.3.1. Le symbole émis le long de cette transition correspond à la partie « non-recouverte » du cosegment correspondant à l'état d'arrivée. Ainsi, sur une transition entre l'état *abc*, et l'état *cdef*, le symbole émis est *def*.

L'automate ainsi construit est non-déterministe. Le langage correspondant à toutes les chaînes d'appariements qu'il est possible de produire en imposant des contraintes de recouvrement est donc lui aussi régulier. Le classement des multiples analyses d'un mot dans cette grammaire régulière est effectué selon la définition *a priori* d'un critère structurel, qui favorise l'utilisation de transitions correspondants à de larges recouvrements.

Tous ces modèles réalisent donc des formes particulières *d'instanciation* de grammaires régulières sur des alphabets de segments. Dans ce contexte, nous l'avons vu, l'utilisation d'un dictionnaire de segments relativement contraint, et d'un critère structurel d'ordonnancement des analyses multiples d'une lexie s'est révélée être la stratégie la plus favorable, cet avantage se traduisant d'une part par des scores de transcription sensiblement supérieurs, et d'autre part par une meilleure utilisation du jeu d'exemples, puisqu'une observation même exceptionnelle est susceptible d'être mise à contribution.

La reformulation que nous venons de proposer pour les algorithmes « analogiques » permet de mieux justifier les réticences que nous avons pu manifester, en particulier envers SMPA et PRONOUNCE, envers l'utilisation d'un tel qualificatif. Quelque soit l'acception que l'on donne à ce terme, et y compris dans son sens le plus faible, SMPA pas plus que PRONOUNCE n'extraient des analogies. Reposant intégralement sur la manipulation de segments, ces algorithmes n'identifient en effet à aucun moment des similarités entre lexies, et sont bien en peine d'identifier les analogues connus d'un terme inconnu, à partir de la prononciations desquels le mot inconnu pourrait être transcrits. En d'autres termes, ils ne fournissent aucune définition formelle à des notions telles que celle de voisinage lexical, définition qui serait pourtant indispensable à des études plus poussées de leur pertinence cognitive (comme par exemple celle qui consisterait à prédire les éléments lexicaux les plus influents, ou à mettre en évidence des effets de fréquence dans la construction des analogies). Leur capacité à reproduire, avec une certaine fidélité, des transcriptions correspondant à la lecture, par l'humain, de pseudo-mots ne doit pas nous leurrer : ces modélisations ne sont pas, à nos yeux, appropriées pour décrire les mécanismes cognitifs par lesquels l'humain parvient à s'inspirer de mots connus pour prononcer des mots inconnus, pas plus qu'elles ne sont adaptées pour réaliser le type de mécanisme de transcription que nous nous fixions comme objectif au début de cette étude. Plutôt que persévérer dans cette direction de recherche, nous avons donc préféré nous tourner vers d'autres développements, qui font l'objet du chapitre suivant.

Il n'en reste pas moins, et les performances de SMPA en sont la preuve, que le cadre très général que nous venons de présenter ouvre des perspectives de recherche intéressantes et potentiellement fructueuses. Il est en effet possible d'envisager avec optimisme un certain nombre d'extensions naturelles dans lesquelles

les aspects positifs des différentes manières d'instancier des grammaires seraient combinés. En particulier, il est envisageable de stochastiser la grammaire régulière de co-segments définie par SMPA, et de remplacer le critère de recouvrement maximal par un critère induit par l'assignation de probabilités aux transitions de l'automate. Ceci conduit à abandonner un principe à nos yeux important, celui de l'utilisation d'une mesure structurelle pour ordonner les recombinaisons licites. L'expérience du modèle co-multigramme a toutefois mis en évidence le fait que les segmentations conjointes utilisant des longs segments étaient, dans un cadre probabiliste, « naturellement » privilégiées. Ce dernier point nous incline à penser que, dans un tel cadre, l'arbitrage entre longueur et fréquence des appariements pourrait, sans que la définition d'un critère *ad hoc* soit nécessaire, favoriser les recombinaisons tirant parti de longs recouvrements. Une telle extension fournirait par ailleurs un moyen direct de lever la condition d'indépendance entre les segments recombinaisonnés, dont nous avons vu qu'elle nuisait aux performances du modèle de multigrammes à deux niveaux. Plus généralement, ce cadre permet d'exprimer de manière relativement naturelle l'interaction entre une modélisation probabiliste et des contraintes symboliques (la contrainte de recouvrement, mais on pourrait en imaginer bien d'autres) dans des modèles de transduction de chaînes, modèles qu'il semble possible d'appliquer à d'autres d'applications mettant en jeu de tels mécanismes (la synthèse de la parole par recombinaison d'unités de taille variable, par exemple).

Chapitre 4

Analogies et paradigmes

4.1 Préambule

4.1.1 Introduction

Nous avons présenté au chapitre précédent un algorithme, JUPA, qui fonctionne en recollant deux segments d'un mot. Nous avons en particulier montré, à la section 3.2.5, que cette technique simple devait une bonne partie de sa réussite à l'utilisation, parfois chanceuse, de relations lexicales que nous avons qualifiées de « paradigmatiques ». À cette occasion, nous avons également mentionné certains des obstacles que cette approche rencontrait dans l'exploitation de ces relations (pseudo-)paradigmatiques. Nous avons par ailleurs constaté, en discutant le modèle des multigrammes à deux niveaux (à la section 3.4), que la prise en compte simultanée des formes graphiques et phonétiques associées à une entrée lexicale permettait, dans une certaine mesure, d'extraire automatiquement des (co)-segmentations morphologiquement pertinentes de ces entrées.

Dans ce chapitre, nous poursuivons cette ligne de recherche en explorant une voie d'extension de JUPA alternative à celle développée aux sections 3.3 et 3.4 du chapitre précédent. Notre but est ici de chercher des moyens systématiques pour utiliser sciemment, dans les mécanismes de transcription par analogie, les relations paradigmatiques qui structurent les lexiques. Nous espérons ainsi être à même de produire des analogies structurelles linguistiquement fondées, reposant en particulier sur l'existence de paradigmes dérivationnels, qui nous permettront d'améliorer les performances de nos systèmes de transcription.

Nous nous inscrivons ici clairement, quoiqu'avec des méthodes sensiblement différentes, dans la ligne de recherche initiée par Stefano Federici et Vito Pirrelli, telle qu'elle est décrite dans (Pirrelli & Federici 1992; 1994b; 1994a; Federici & Pirrelli 1994b; 1994a). Notre approche se distingue toutefois de la leur sous trois aspects importants :

- à une approche « diffuse » de l'analogie fondée sur des patrons d'activités distribués au travers de liens d'activation/inhibition dans un réseau de

symboles, nous plaçons pour une approche plus formelle de l'analogie, exprimée dans les termes d'un rapport de proportion entre quatre termes, qui permet en particulier de donner un sens précis au concept d'analogie;

- à une approche strictement fréquentielle de l'apprentissage, par laquelle les liens les mieux attestés sont renforcés, nous plaçons, dans la stricte continuation des approches développées dans le chapitre 3, pour une approche qui ordonne prioritairement les analogues sur la base d'un critère structurel;
- à une approche de l'apprentissage vue comme méthode d'extraction de correspondances entre des éléments dans deux domaines (langages) distincts, nous plaiderons pour des techniques d'acquisition de couplages de transformations internes à chacun des domaines.

Nous avons organisé ce chapitre de la manière suivante : après un exemple introductif qui nous permettra de donner l'intuition générale de notre démarche, nous présentons, à la section 4.2, une introduction sur le concept linguistique de paradigme, en relation tout particulièrement avec, d'une part, les théories morpho-phonologiques du lexique, et d'autre part, avec les théories de l'analogie telles qu'elles sont mises en œuvre dans divers domaines de l'intelligence artificielle. Cette discussion sera principalement pour nous l'occasion de clarifier ce concept, et d'avancer un certain nombre de faits qui justifient une approche paradigmatique (au sens large) des phénomènes morpho-phonologiques.

Aux sections suivantes (4.3 et 4.4), nous présentons en détail deux algorithmes de transcription fondés sur le concept d'*exploration paradigmatiquement contrainte d'un lexique*, algorithmes qui, d'une certaine manière, font pendant à JUPA et SMPA : le premier ne met à contribution que deux segments extraits de lexies connues, et est généralisé par le second, qui utilise potentiellement un nombre quelconque de segments. Comme dans les chapitres précédents, une large place sera faite à l'évaluation comparative de ces deux nouvelles stratégies. Nous terminerons ce chapitre en appliquant, à la section 4.5, l'approche paradigmatique à un autre problème d'apprentissage du langage naturel abondamment débattu, l'apprentissage de la conjugaison des verbes anglais.

Notre contribution entièrement personnelle à ce chapitre concerne principalement les sections 4.4 et 4.5, le modèle présenté dans la section 4.3 ayant été conçu en collaboration avec Stefano Federici et Vito Pirrelli, mais mis en œuvre et évalué par nos soins.

4.1.2 Retour sur l'analogie *Martin/Martine*

L'idée de mettre en œuvre une stratégie paradigmatique trouve son origine dans l'examen détaillé d'une erreur surprenante de la procédure de recollement implantée par JUPA : le prénom *Martin* est incorrectement prononcé /martin/. En examinant les deux lexies impliquées dans cette transcription erronée, on s'aperçoit que si l'une, *Martine*, est un bon candidat à l'analogie, l'autre, *Amin* (qui est de fait responsable de l'incorrection de la terminaison) est bien moins pertinent : c'est un nom importé, dont l'orthographe comme la prononciation

sont assez flottantes, et qui s'oppose à toute une classe de prénoms depuis longtemps francisés : *Augustin*, *Constantin*, *Albertin*, etc. Le recollement (/martin/ + /amin/) des transcriptions de ces deux mots est néanmoins celui qui fournit le plus grand niveau de recouvrement (les deux derniers phonèmes des deux chaînes sont identiques), d'où la transcription /martin/.

Ce qui choque dans cette erreur, c'est que la procédure de recollement se laisse pour ainsi dire flouer par la pression analogique exercée par un item exceptionnel du lexique des noms propres français, en l'occurrence *Amin*, et passe ainsi à côté d'une alternance régulière massivement attestée dans notre lexique, à savoir l'alternance /ē/-/in/ que l'on retrouve dans les paires *Constantin-Constantine*, *Perrin-Perrine*, ainsi que dans des dizaines d'autres couples.

L'utilisation d'un critère fréquentiel, remède « naturel » contre l'influence des exceptions, et qui devrait permettre de filtrer la prononciation incorrecte, se révèle, dans ce cas précis, particulièrement inopérante : *Amin* est loin d'être isolé dans notre lexique, où il côtoie *Azzedin* et *Aydin* ou encore *Erwin* et *Marvin*. En conséquence, de très nombreux recollements confirment la transcription /martin/, ce qui contribue à dissimuler le caractère profondément irrégulier de cette forme.

La solution à ce type d'erreur, qui est loin d'être isolée (c.f. la discussion de la section 3.2.5), passe à notre sens par la prise en compte explicite de ce qui distingue essentiellement la première classe d'analogues (celle comprenant *Constantin*, *Marin*, etc.) de la seconde (*Amin*, *Kevin*, etc.), à savoir le fait que les premiers participent à un système de paires alternantes (schématisé à la figure 4.1 par un *graphe de commutation*), comme en attestent les formes *Constantine* et *Marine*, alors que ni *Amine* ni *Marvine* ne sont observées.

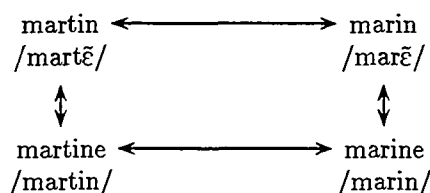
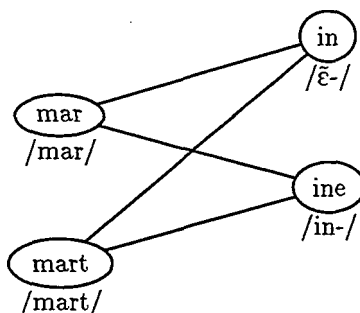
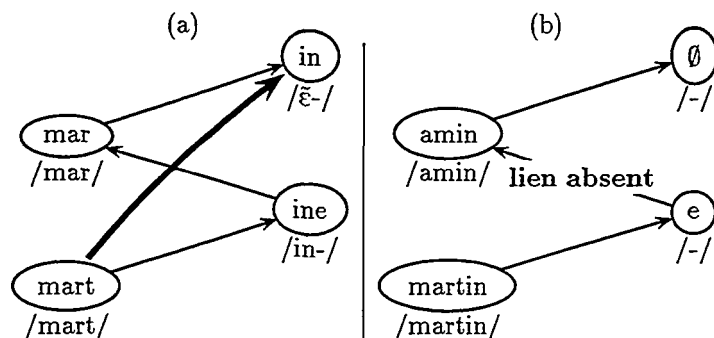


FIG. 4.1 – Un système d'alternances

Une autre présentation de ce type de schéma alternant est donnée dans les travaux de Pirelli et Federici (1994b; 1994a) sous la forme du graphe représenté à la figure 4.2.



FIG. 4.3 – *Extension paradigmatique*FIG. 4.2 – *Un réseau d'alternances*

Ce que cette représentation met en évidence, c'est l'existence d'un segment initial (*mart*), commun aux formes (graphiques et phonétiques) du couple *Martine-Martin*, qui « fonctionne » de manière homologue au segment initial *mar* partagé par les deux formes du couple *Marine-Marin*; l'homologie signifiant ici que ces deux segments supportent de la même façon (i.e. avec les mêmes conséquences sur la forme phonologique produite) la concaténation des segments finaux *ine* et *in*. De manière duale, la même remarque vaut pour les segments terminaux *ine* et *in*, dont la prononciation est indépendante du préfixe (*mart* ou *mar*) auquel ils sont rattachés. Nous désignerons dans la suite de la présentation les segments correspondant aux nœuds du graphe de la figure 4.2 sous le terme de *noyaux paradigmatiques*, et les liens qui les relient seront appelés *liens paradigmatiques*. Une hypothèse forte de notre démarche sera alors que ces noyaux paradigmatiques correspondent à des segments qui sont phonologiquement stables, i.e. qui vont se prononcer d'une manière identique indépendamment des segments auxquels ils sont accolés. Federici et Pirrelli (1994b) ont montré que l'extraction à partir d'un lexique de tels noyaux paradigmatiques était algorithmiquement faisable (quoique très coûteuse) et qu'elle représentait un moyen approprié pour retrouver des paradigmes morphologiques.

Pour montrer la pertinence de ce type d'approche, replaçons-nous dans la situation initiale, à savoir celle où le mot *Martin* est inconnu, où sa prononciation doit être inférée, et considérons les deux graphiques 4.3a et 4.3b.

La différence entre les deux classes d'analogues pour le mot *Martin* s'exprime au travers de ces schémas de manière simple : pour les mots de type *Marin*, il existe un chemin de liens paradigmatiques (de *mart* à *ine*, de *ine* à *mar*, de *mar* à *in*) qui permet de relier le segment *mart* au segment *in*. Une extension paradigmatique (figurée par l'arc reliant directement *mart* à *in*) de ce chemin est alors possible, permettant de dériver par concaténation une prononciation (correcte) /martɛ-/.

Dans le cas des éléments de l'autre classe d'analogues, un tel parcours est impossible. Il manque en effet un lien entre les segments *Amin* et *e*, lien dont l'absence résulte précisément de la non-existence de la forme *Amine*. En consé-

quence, l'extension paradigmatique est bloquée; *Amin* ne peut être utilisé pour inférer une transcription de *Martin*, et la seule construction possible est celle schématisée sur la figure 4.3a.

Une solution au problème posé par des couples de la forme *Martin-Martine* à des mécanismes de recollement de segments est donc de contraindre les recollements en imposant des conditions de cheminement paradigmatique. Pour l'essentiel, les techniques développées dans ce chapitre visent à formaliser cette intuition première.

4.2 Morphologie, Analogie, Phonologie

Nous avons, à la section précédente, introduit de façon intuitive le qualificatif de *paradigmatique*, pour désigner un système régulier d'alternances. En reconsidérant l'exemple original de la figure 4.2, on s'aperçoit qu'au niveau phonologique deux types d'alternances sont mises en jeu : d'une part celle des préfixes /mart/ et /mar/, qui permet de d'opposer les deux prénoms *Martin* et *Marin*, alternance que l'on peut qualifier de strictement contingente; d'autre part l'alternance entre /ɛ/ et /in/, qui elle est partiellement motivée, dans la mesure où elle reproduit un schéma d'opposition entre féminin et masculin qui se retrouve à de multiples endroits du système flexionnel du français.

Du premier type d'alternances, nous avons, pour l'instant, fort peu à dire. Nous les voyons comme un des effets inéluctables du système d'écriture alphabétique, qui n'utilise qu'un petit nombre de symboles pour représenter un grand nombre de mots, et qui est, de plus, soumis à une forte pression diachronique vers la réduction et la simplification de la forme des mots les plus fréquemment utilisés. Que dans ce contexte certains voisinages lexicaux soient spécialement peuplés, d'autant plus peuplés que les mots sont courts, ne saurait nous étonner.

Le second type d'alternance, qui s'interprète comme une relation morphologique, nous intéresse bien plus. Nous souhaitons donc, dans cette section, replacer le terme de *paradigme* dans son contexte originel, celui de l'analyse morphologique des systèmes flexionnels; et justifier, sur un plan linguistique, l'approche de l'analogie qui sera mise en œuvre dans les sections 4.3 et 4.4.

Le point précis sur lequel nous voulons tout spécialement insister est qu'un lexique de taille raisonnablement grande n'est pas une simple liste informe de formes sans rapport entre elles; mais qu'au contraire cette liste présente une structure, un mode d'organisation particulièrement riche. En examinant quelques-unes des descriptions linguistiques de cette structure, nous suggérerons qu'il est possible d'en capturer une bonne partie par le biais du concept de *paradigme*, que nous mettrons en relation avec le concept d'*analogie*, et montrerons qu'une telle analyse se justifie également lorsqu'on considère des lexiques de noms propres.

4.2.1 Paradigmes en morphologie

Nous présentons ici une brève revue de l'emploi du concept de paradigme dans la description linguistique des processus morphologiques. Nous présenterons d'abord la genèse de ce concept, en relation avec la description des systèmes inflexionnels. Nous mentionnerons ensuite quelques-unes des critiques qui ont pu lui être adressées, avant de signaler quelques phénomènes pour lesquels une analyse en termes d'alternance régulière reste justifiée. Pour l'essentiel, les idées développées dans cette section s'inspirent la très complète étude des théories génératives de la morphologie présentée dans (Spencer 1991).

Paradigmes flexionnels en morphologie

La notion de *paradigme* prend son sens premier en morphologie dans l'étude des systèmes *flexionnels*, où elle désigne « the set of all the inflected forms which an individual word assumes. » (Spencer 1991, p11). On retrouve dans cette définition l'intuition essentielle, celle de *classe d'éléments alternants*; intuition qui est exemplairement assignée à ce terme dans la tradition structuraliste, qui conçoit la dimension paradigmaticque comme la dimension selon laquelle s'organisent les unités qui sont susceptibles de se remplacer mutuellement dans un même syntagme. Un ensemble paradigmaticque est donc essentiellement un ensemble d'oppositions (Martinet 1967, I-20). Par extension, on désigne sous ce même vocable l'ensemble des radicaux qui suivent des schémas de flexionnels similaires (par exemple les verbes du premier groupe en français).

La forme archétypale d'un paradigme flexionnel est celle d'un ensemble de formes construites sur un radical commun, et où chaque ensemble de traits syntaxiquement pertinent s'exprime par l'agglutination d'un morphème spécifique. C'est par exemple le cas de certains adjectifs en français, (ainsi *grand*), pour lesquelles chacune des oppositions syntaxique (genre et nombre) s'exprime par l'ajout d'un morphème, conduisant au paradigme complet¹ : {*grand, grande, grands, grandes*}. Le finnois et le hongrois sont deux exemples de langues dont le système de flexion est assez proche de ce type d'organisation idéale, les langues indo-européennes présentant des formes plus ou moins dégénérées d'une telle organisation.

La ressemblance formelle entre un tel système d'organisation des flexions, et le concept d'analogie (ou d'*homologie*), envisagé sous l'angle d'un rapport de proportion, a été très tôt perçue (Hoffman 1995, p 15) par les linguistes. Dans la tradition structuraliste, diverses tentatives de formalisation mathématique des systèmes d'oppositions ont donc ainsi entreprises, visant par exemple à définir des mesures formelles permettant de mesurer la plus ou moins grande conformité du système flexionnel d'une langue vis-à-vis d'un tel mode d'organisation (Marcus 1967).

Sans entrer dans le détail de ces tentatives, retenons-en que ce qui caractérise en premier lieu l'organisation paradigmaticque d'un système d'inflexions,

1. La description (notamment générativiste) des mécanismes morphologiques tend à ne considérer que les modifications qu'ils induisent sur les représentations phonologiques. Il serait donc plus orthodoxe, mais sans doute moins lisible, de considérer la liste : {/gʁɑ̃d/, /gʁɑ̃d+ə/, /gʁɑ̃d+z/, /gʁɑ̃d+ə+z/}.

c'est l'existence d'un stock relativement petit d'affixes qui sont susceptibles de s'accoler à un nombre beaucoup plus grand de radicaux, pour altérer *de façon systématique* les traits syntaxiques dont les radicaux sont porteurs.

Limites de la description paradigmatique

L'observation d'un certain nombre de déviations par rapport à ce type d'organisation morphologique a toutefois conduit à remettre en question la pertinence du concept de paradigme, en relation avec l'universalité des mécanismes d'affixation. La description des systèmes flexionnels d'un bon nombre de langues n'est en effet envisageable sous la forme de tels mécanismes que si l'on fait l'hypothèse que les affixes sont en fait des unités particulièrement abstraites, qui ne sont pas directement observables dans les formes de surface. Sans qu'il soit besoin de faire appel à des exemples particulièrement complexes de systèmes flexionnels, mettant en jeu des phénomènes de reduplication, de gémination, ou d'alternance vocalique, mentionnons simplement quelques unes des déviations les plus flagrantes qui sont manifestes dans les langues sur lesquelles nous avons été conduits à travailler, et qui mettent en évidence :

- des phénomènes d'*allomorphie*, reflétés par une multiplication des formes attestées pour un morphème donné. Dans les cas favorables, cette allomorphie est strictement conditionnée par des règles phonologiques; dans les cas moins favorables, elle ressort d'autres causes (étymologique, etc.); dans tous les cas, néanmoins, elle contribue à l'obscurcissement de l'organisation paradigmatique. Un exemple d'allomorphie phonologiquement conditionnée est celui de la flexion verbale marquant la forme passée des verbes anglais réguliers (voir pour une description plus complète la section 4.5) : on observe en effet en surface trois réalisations différentes du morphème flexionnel, à savoir /t/ (dans la conjugaison de *pass*), /d/ (dans la conjugaison de *tag*), et /ɪd/ (dans la conjugaison de *need*);
- des phénomènes de *fusion*, conduisant un morphème unique à porter indissociablement plusieurs traits syntaxiques : ainsi en français *leur* porte de manière inséparable les traits *possessif* et *pluriel*;
- des phénomènes de *synchrétisme*, conduisant une forme de surface à représenter simultanément suivant les contextes plusieurs ensembles différents de traits syntaxiques (le système de conjugaison du français en contient de nombreux exemples, comme en atteste la confusion des formes de l'indicatif présent et du subjonctif présent des verbes du premier groupe).

Si les systèmes flexionnels peinent, en conséquence, à être décrits de manière satisfaisante dans le cadre d'oppositions systématiques des formes de surfaces, la situation est sans doute aggravée lorsque l'on envisage les mécanismes dérivationnels. Ceux-ci présentent un certain nombre de différences importantes avec les mécanismes flexionnels, différences qui, d'une certaine manière, rendent encore moins pertinente l'utilisation du concept de paradigme pour décrire les

familles de dérivés. En particulier :

- les inflexions s'appliquent uniformément, y compris sur des formes dérivées. Ceci entraîne deux conséquences importantes. D'une part, étant donné un ensemble de traits morpho-syntaxiques, et une forme radicale susceptible de porter ces traits, il existe (presque) toujours un dérivé flexionnel de la forme radicale les portant. Ainsi, en français, si l'on omet les rares défectifs, tous les verbes possèdent une forme correspondant à la troisième personne du pluriel de l'imparfait. À l'inverse, certains verbes ne possèdent pas de dérivé correspondant à la nominalisation pour désigner l'action correspondant au verbe : ainsi *manger*, à notre connaissance, n'en possède pas (ni *mangeage*, ni *mangement*, ni *mangeation* ne sont attestés, et tous nous paraissent assez douteux);
- les formations dérivationnelles participent à la dynamique du lexique d'une langue : un dérivé régulier, originellement porteur d'une signification complètement co-déterminée par la lexie à laquelle il se rattache et la dérivation qui le produit, peut progressivement acquérir une autonomie lexicale, fixer des significations idiosyncrasiques propres, et se libérer au court du temps sa stricte dépendance sémantique par rapport au terme d'origine. Les exemples de ce phénomène abondent, et concernent de la même manière les mots composés. Ceci ne s'observe pas des dérivés flexionnels, qui sont sémantiquement transparents.

L'ensemble de ces observations a donc conduit à la remise en cause le pouvoir explicatif de ce concept dans le cadre d'une description unifiée et cohérente des mécanismes morphologiques. Les approches générativistes de la morphologie tendent en conséquence à ignorer le concept de paradigme, ne voyant dans ce mode d'organisation du lexique qu'un épiphénomène de processus dont l'intuition est intégralement capturée par le concept de règles de composition et de dérivation. C'est dans cette direction qu'ont eu lieu les développements les plus récents en morphologie, conduisant d'une part à l'établissement de grammaires morphématiques (dans laquelle les morphèmes sont des unités très abstraites), et d'autre part à une définition précise de l'interfaçage entre ces règles de grammaire et le module de traitement phonologique.

Sans entrer dans les nombreux débats théoriques qui agitent ce courant de réflexion, ni dresser un réquisitoire en règle contre une telle conception de la morphologie (voir par exemple (Bybee 1988)), nous voudrions signaler deux phénomènes importants sur lesquels la modélisation générativiste semble buter, et pour lesquels une analyse en termes d'alternances semble conserver tout son intérêt, qui sont les paradoxes de parenthésage, et les dérivations « approximatives ».

Formations analogiques

Les paradoxes de parenthésage Un phénomène particulièrement résistant aux analyses fondées sur la construction de structures hiérarchiques de morphèmes, connu sous le nom de *paradoxe de parenthésage*, a toutefois conduit

à remettre en lumière le concept de paradigme comme un concept pertinent pour la description des relations lexicales (sur cette question, nous renvoyons à nouveau à (Spencer 1991), tout particulièrement au chapitre 11, qui fournit une longue liste de paradoxes, et introduit quelques unes des hypothèses linguistiques permettant de justifier leur existence).

Dans leur définition la plus générale, les paradoxes de parenthésage se manifestent par l'existence de deux parenthésages différents d'une forme donnée, chaque parenthésage référant à la description d'un fait linguistique particulier. Ces paradoxes portent sur des aspects très variés des constructions morphologiques, mais tous posent d'une certaine manière le problème des unités théoriques manipulées dans les descriptions lexicales.

Un premier exemple, abondamment discuté (parce que remettant en cause une hypothèse théorique importante, celle d'un ordre rigide dans l'application des dérivation), est celui de *ungrammaticality*: les conditions de restriction morpho-syntaxiques sur la préfixation de *un* imposent que ce morphème modifie un adjectif, suggérant le schéma dérivationnel suivant :

grammatical $\xrightarrow{Adj \rightarrow Adj}$ *ungrammatical* $\xrightarrow{Adj \rightarrow N}$ *ungrammaticality*;

et le parenthésage : (((*un*)(*grammatical*))(*ity*)). L'hypothèse théorique selon laquelle il existerait en anglais deux classes d'affixes distinctes, hypothèse motivée par l'observation de deux types de comportements morpho-phonologiques très différents² suggère un parenthésage différent. En effet, si l'on suit cette hypothèse, alors *ity* doit intervenir dans la dérivation avant *un*, et le parenthésage correct est : ((*un*)((*grammatical*)(*ity*))).

Si ce paradoxe-là n'est troublant que dans le cadre d'une théorie donnée (celle qui suppose un ordre rigide des affixations), le même type de contradiction apparaît, à un niveau différent toutefois, dans des constructions telles que *sous-marinier* ou *macro-économiste*, et dans lesquelles le contenu sémantique (le *sous-marinier* n'est pas inférieur au *marinier*, ni le *macro-économiste* nécessairement plus gros que l'*économiste*) impose un parenthésage qui diffère du parenthésage typographique. Plus subversifs encore vis-vis d'une analyse morphologique par succession de dérivation sont des exemples dans lesquels un terme dérivé ne peut s'analyser que par un mécanisme de *désaffixation*, comme le composé *theoretical linguist*, qui ne peut se concevoir que comme un dérivé de *theoretical linguistics* (un exemple voisin en français est *peintre abstrait*, qui hérite sémantiquement de l'idiosyncrasie sémantique cristallisée dans le composé *peinture abstraite*).

Une solution élégante à ce problème (Spencer 1988) est de postuler l'existence d'un mécanisme autonome de production lexicale, uniquement fondé sur le concept de *rapport de proportion*. En dernière instance, le *macro-économiste*

2. Les affixes de classe I (*-ity*, *in-*, *-ic*, etc) sont, selon cette théorie, moins productifs, moins transparents sémantiquement, et entraînent des réorganisations phonologiques importantes (changements vocaliques, assimilations, déplacement de l'accent). Ceci n'est pas vrai des suffixes de la classe II (*-hood*, *-ness*, *un-*, etc.). À titre d'exemple, il est instructif de considérer la différence de comportement entre *in* et *un*, dans des formes comme *in+legal* → *illegal* (assimilation du /n/); *un+like* → *unlike* (conservation du /n/). Pour une discussion plus complète de la validité de cette distinction, voir par exemple (Durand 1990, pp 170-172).

ne s'analyse que comme le complément le plus adéquat de la série *économie-économiste-macro-économie*, établissant un double (en fait un triple) rapport de proportion entre les termes de cette série, rapport que l'on peut gloser en « le macro-économiste est à la macro-économie ce que l'économiste est à l'économie », et qui est également valide dans le domaine des représentations graphiques et phonétiques.

Pression analogique et morphologie approximative Une seconde difficulté que rencontre une approche strictement générativiste des phénomènes morphologiques dérivationnels est liée à l'existence de lexies résultant d'une productivité autonome de relations paradigmatiques, lexies qui sont en conséquence difficilement analysables en termes de suffixation.

Un exemple caractéristique de ce type de production est le néologisme anglais *workholic*, construit en dépit de l'existence du suffixe « naturel » *ic*, sur le strict modèle d'*alcoholic*, dont la terminaison a été extraite sur des bases prosodiques plutôt que morphologiques. On doit à Saussure (1916, chap. V) la première analyse de la productivité diachronique de ces fausses décompositions. Considérons un autre exemple, fournit par cet auteur : le verbe *somnoler*, est construit à partir de *somnolent*; toutefois la racine extraite lors de la construction (*somnol*) est étymologiquement incorrecte, puisque *somnolent* se décompose *somn-olent* (litt. « assoiffé de sommeil »). Un grand nombre de formes ne peuvent en fait s'analyser que dans ces termes. L'observation de ce type de formations, conduit Saussure à affirmer le rôle moteur des productions analogiques (i.e. qui s'expriment dans les termes d'un rapport paradigmatique) dans l'évolution des langages. En particulier, ces productions, qui vont toujours dans le sens de la régularisation, contre-balancent diachroniquement les effets de l'érosion phonétique, et permettent de sauvegarder en partie la systématité des alternations et des « relations associatives ».

Notons finalement que bien des exemples de ces productions déviantes sont attestés dans les « bas-côtés » du langage officiel, en particulier dans les néologismes argotiques où le principe de ressemblance analogique joue à plein : ainsi *parigot* où le pseudo-suffixe est extrait de *gigot* (François 1968).

4.2.2 Le cas des noms propres

Aussi paradoxal que ce fait puisse paraître, l'organisation du lexique des noms propres du français (la remarque vaut sans nul doute pour d'autres langues voisines) présente sous de nombreux aspects le même principe d'organisation paradigmatique. Comme nous le verrons dans les paragraphes qui suivent, ce mode de structuration du lexique patronymique peut s'analyser comme résultant d'une part de l'utilisation massive d'un petit nombre de schémas de construction des noms propres, schémas qui ont utilisés uniformément sur tout le territoire pendant les siècles durant lesquels, du bas moyen-âge à la renaissance, l'usage d'un système de double nomination s'est progressivement enraciné; et d'autre part de la forte diversité dialectale, durant cette période, du français. Après avoir décrit sommairement ces deux phénomènes et leur interaction, nous montrerons que dans une large mesure, le même type d'organisation se retrouve dans les

nombreux noms propres d'origine étrangère aujourd'hui intégrés dans le stock de noms français.

Quelques remarques sur le lexique français

Pour l'essentiel, ce court panorama de la genèse du lexique de patronymes français suit la présentation très complète de (Dauzat 1949).

Dérivations hypocoristiques Les noms propres d'origine française se classent en quatre groupes principaux : ceux qui sont construits à partir d'un prénom (e.g. *Bernardin*), ceux qui réfèrent à un lieu d'habitation ou d'origine ethnique (e.g. *Dumoulin*, *Poitevin*), ceux qui renvoient à des noms de métiers (e.g. *Charpentier*), et ceux qui émanent à l'origine d'un sobriquet (e.g. *Legrand*).

Indépendamment de leur origine, mais ce besoin s'est surtout fait sentir, pour des raisons manifestes liés à leur concentration, pour les prénoms, les noms originaux ont donnés naissance à de nombreux diminutifs, prenant la forme de dérivation par suffixation (e.g. *Jean* → *Jeannot*, *Jeannin*, mais aussi *Boucher* → *Boucherot*, *Blond* → *Blondin*); ou plus rarement par aphérèse (*Nicolas* → *Colas*, *Colin*...).

Des causes différentes pouvant produire les mêmes effets, l'utilisation d'un système de cas (cas-régime de l'ancien français, ou génitif emprunté au latin) a donné plus rarement naissance à des paradigmes « flexionnels » semblables en tous points aux séries dérivationnelles; ainsi *Guy* → *Guyon* (cas-régime, on notera que le suffixe *on* existe indépendamment comme suffixe hypocoristique), ou encore *Fabre* → *Fabri* ou encore *Fabry* (par imitation du génitif latin).

Erreurs et confusions Un second facteur important s'est conjugué à ces phénomènes strictement morphologiques, que l'on peut désigner sous le terme de *flottement généralisé* de l'orthographe des noms propres. Sous ce générique, on regroupe toutes les erreurs d'interprétations ou tout simplement de recopiage; comme les fausses étymologies et les analogies hasardeuses qui, en l'absence d'une institution capable de réguler ces puissants facteurs de changement diachronique, ont conduit à la multiplication des orthographes possibles d'un nom donné.

Ainsi, le suffixe germanique *ald* a évolué en conformité avec le reste du lexique, se transformant par amuïssement du *l* en *aud*, voire en *aut*; la restauration malheureuse du *l* donnant naissance à de nouvelles formes *ault* et *auld*, soit au final pour chacun des prénoms en *ald* au moins cinq formes dérivées. D'une manière plus générale, l'examen d'un large corpus nous a largement convaincu que chaque fois qu'un phonème ou une suite de phonèmes peut s'orthographier de plusieurs manières, toutes (ou presque) sont attestées. Parmi les cas les plus marquants de ces confusions, mentionnons la présence de nombreux *h* ou de *s* incongrus (*Thévenin* à côté de *Tévenin*, hypocoristique d'*Étienne*; *Besnard* à côté de *Bénard* (qui dérive en fait de *Bernard*, par amuïssement du *r*), de consonnes « inutilement » répétées (*Collin*, *Collet* ...), etc.

Jean	Guillaume	Pierre
Jeanne	Guillot	Perrin
Jeannin	Guillou	Perrier
Jeannot	Guillet	Perrot
Jeanneau	Guillon	Perret
Jeannet	Guillemin	Perron
Jeanneret	Guillard	Perraud
Jeanney	Guillemot	Perrault
Jeannetot	Guilloux	Perreau
Jeannerot	Guillemet	Perraudin
Jeannette	Guillaud	Perry
Jeanniot	Guilloteau	Perrotin
Jeannel	Guillotin	Perroud
Jeannesson	Guillemain	Perrichon
Jeanniard	Guillaumin	Perroux

TAB. 4.1 – Quelques paradigmes patronymiques du lexique français.

Au bout du compte Ce que l'on observe d'un point de vue synchronique, c'est que la conjonction des mécanismes de composition, les variations régionales des suffixes hypocoristiques ((e)quin dans le Nord, *esson* ou *eçon* dans l'est, *aud* dans le Sud-Ouest ...) et les déformations orthographiques, ont conjointement contribué à la formation de classes qui présentent en surface les caractéristiques d'une organisation paradigmatique. Les courts extraits présentés dans table 4.1 des séries correspondant aux noms de baptême *Jean*, *Guillaume* et *Pierre*, et dont tous les membres sont fortement attestés (leur fréquence est supérieure à 100 dans le lexique de noms propres dont nous disposons), donnent une illustration de ce phénomène.

Cette observation justifie ainsi *a priori* l'utilisation pour les noms propres des mêmes techniques analogiques que celles utilisées pour traiter des lexiques de noms communs.

Les noms étrangers

Dérivations Nous l'avons déjà noté (à la section 1.1.2), le succès d'un système de transcription des noms propres dépend crucialement du traitement correct des éléments importés. Nous pensons qu'ici encore, l'approche analogique fondée sur l'extraction de paradigmes se justifie pleinement, et ce à double titre. Tout d'abord parce que de nombreuses langues exhibent des mécanismes de construction des noms propres par (pseudo) dérivation très similaires à ceux énumérés pour le français; le plus connu étant sans nul doute la dérivation patronymique, marquant par un suffixe le lien de filiation, et qui est attesté aussi bien dans les langues germaniques (sous la forme des suffixes *son* en anglais, *sohn* en allemand, *sen* en danois), slaves (*ovitch* en russe), que romanes (*escu* en roumain)... Un autre mécanisme conduisant aux mêmes résultats est l'utilisation par certaines langues (exemplairement les langues slaves) d'un système de flexion en genre ou en cas applicable y compris aux noms propres : ainsi le féminin tchèque *ova*, utilisé pour dériver du nom du mari (ou du père) le nom de la femme (ou de

la fille). Inscrits sur les papiers des immigrant(e)s, ces formes fléchies ont perdu leur interprétation grammaticale, et se sont fixées telles quelles dans le lexique français, où elles figurent bien souvent au côté des formes radicales d'origine.

Problèmes orthographiques Ensuite parce que les flottements orthographiques attestés pour les orthographes françaises trouvent leur équivalent à une échelle encore plus grande pour les noms empruntés. D'une manière générale, ce flottement exprime un conflit entre la volonté de la conservation de la forme graphique originale, et celle de garantir une prononciation approximativement correcte. Un bon exemple de ce phénomène est fourni par les noms d'origine italienne contenant un *e* graphique intérieur, pour lesquels on observe également systématiquement la variante présentant un *é* graphique. Un autre exemple de ce type de flottement concerne la transcription de noms d'origine polonaise se terminant en /itʃ/, suivant que l'orthographe française imite la graphie originale (et maintient le *cz* original), ou préserve une prononciation approximative (son remplacement par *tch*).

Ce phénomène est encore plus marqué pour les noms empruntés à des langues utilisant des systèmes orthographiques complètement différents du nôtre (noms russes, bulgares, et surtout d'Afrique du Nord) : nous avons ainsi, par exemple, observé pas moins d'une quarantaine de transcriptions différentes du nom *Mohamed*, correspondant à des fluctuations sur le nombre des consonnes, et la position et la nature des voyelles.

L'organisation paradigmatique comme un mécanisme d'identification Enfin parce qu'il nous semble que l'appartenance à une série paradigmatique donnée est un indicateur relativement pertinent de l'origine d'un nom propre, intuition que l'exemple du couple *Martin/Martine* discuté à la section 4.1.2 permet partiellement de saisir. En voulant fonder la prononciation d'un mot sur celle d'un analogue avec lequel il entre dans un rapport d'homologie paradigmatique, nous augmentons nos chances pour que les deux termes de l'analogie aient une origine commune; et qu'en conséquence leurs prononciations présentent finalement des traits communs. Il est important de noter qu'ici encore, c'est parce que nous fondons l'analogie sur un principe structurel plutôt que fréquentiel que nous serons en mesure de capturer des micros-régularités telles que celles qui organisent les groupes de noms propres issus de telle ou telle langue étrangère, et dont les séries paradigmatiques sont nécessairement plus « creuses » que les séries de patronymes autochtones.

4.2.3 En guise de conclusion

Alternances, oppositions, analogies et paradigmes

Indépendamment du débat théorique sur le pouvoir explicatif du concept de paradigme dans le cadre de la description synchronique et diachronique des systèmes morphologiques, débat dont nous avons brièvement rappelé quelques éléments importants, il nous semble nécessaire de conclure cette brève incursion

dans le domaine de la linguistique en remettant en perspective ces éléments dans le contexte de notre travail sur la transcription par analogie.

Le principal enseignement de cette étude, aura été, à nos yeux, d'apporter des éléments justifiant l'utilisation des relations paradigmatiques dans le cadre de l'apprentissage de la transcription graphémo-phonétique. En effet, même si la totalité des mécanismes morphologiques ne se laisse pas analyser à l'aide du concept de paradigme, il n'en reste pas moins que les oppositions morphologiques se manifestent, en surface, par des alternances phonologiques (ou graphémiques, dans le domaine écrit), alternances qui possèdent une certaine régularité et qui manifestent une certaine cohérence interne. C'est précisément à ce type d'alternance de surface que nous référerons dans nos emplois ultérieurs du terme *paradigmatique*.

De plus, nous avons vu que ces systèmes d'oppositions renvoient à la même figure formelle d'un *rapport de proportion* que les modèles de l'analogie, ce qui nous conforte dans leur étude dans le cadre de la construction de modèles de prononciation par analogie.

Finalement, en ce qui concerne plus particulièrement les langues sur lesquelles nous travaillons, ces alternances se manifestent principalement sous la forme de l'affixation d'éléments dont la forme est globalement indépendante de la racine sur laquelle ils se greffent, ce qui justifie notre choix de travailler uniquement sur des relations d'oppositions observables aux extrémités des formes.

Ces trois éléments nous incitent donc à penser qu'étant donné un lexique, il est possible et fondé de s'appuyer sur l'utilisation de telles relations d'alternances entre lexies, pour extraire les « règles » de prononciation non pas à partir d'associations statiques entre les chaînes graphiques et phonétiques, mais à partir de l'observation conjointe des trajectoires dérivationnelles dans les domaines graphiques et phonétiques.

4.2.4 Remarques complémentaires

À ces motivations linguistiques qui incitent à l'étude des systèmes d'alternance, nous voudrions rajouter deux considérations pratiques qui nous confortent dans cette direction. La première est que les systèmes actuels (et, semble-t-il, futurs) de synthèse de la parole vont s'appuyer sur des lexiques de mots transcrits de plus en plus vastes. Dans ce contexte, les seuls mots pour lesquels la transcription ne résultera pas d'un accès dictionnaire direct, et pour lesquels les systèmes de transcription seront réellement sollicités, seront les créations lexicales et les emprunts. Intégrer au principe de fonctionnement du système de transcription lui-même les plus productifs des mécanismes par lesquelles ces créations lexicales apparaissent, à savoir des principes morphologiques, nous semble en conséquence être une manière potentiellement féconde de procéder.

La seconde raison est que l'appartenance à un schéma dérivationnel connu nous semble un critère tout-à-fait pertinent pour identifier avec une bonne précision les éléments inconnus d'un texte qui ont une « bonne allure » de mots du langage, et à l'inverse, ceux qui sont susceptibles d'être des emprunts lexicaux à d'autres langues. Dans la mesure où les emprunts sont, et c'est tout spécia-

lement vrai des noms propres, des mots particulièrement délicats à transcrire, l'approche paradigmatique semble capable d'intégrer dans un même mécanisme un processus d'identification des emprunts et un processus de transcription : elle présente donc, dans le cas de la transcription des noms propres, un double intérêt.

4.3 Prononcer par construction d'analogies paradigmatiques³

Nous présentons dans cette section un premier algorithme exploitant de façon explicite les relations paradigmatiques pouvant être extraites d'un lexique. Cet algorithme est en fait une mise en œuvre directe de l'intuition exposée dans l'introduction de ce chapitre. Afin d'en faciliter la compréhension, nous illustrerons notre démarche au fur et à mesure des besoins par l'exemple, déjà utilisé à la section 4.1.2, du couple *Martin-Martine*.

Dans la mesure où l'intuition générale qui fonde cette procédure a déjà été exposée, nous commençons directement par présenter la formalisation algorithmique, ainsi qu'un certain nombre de remarques complémentaires. Nous donnons ensuite une série de résultats expérimentaux obtenus en appliquant cette méthode au problème de la transcription, avant de discuter ces résultats et de suggérer quelques voies d'amélioration.

4.3.1 Mise en œuvre d'une stratégie paradigmatique

Présentation formelle (avec illustration)

Formellement, la stratégie mise en œuvre pour contraindre des schémas de recollement par un critère d'appartenance à système régulier d'alternances s'exprime de la façon suivante (voir l'algorithme 4). Nous avons baptisé cette procédure PPP, appellation qui sera fréquemment utilisée par la suite.

Pour faciliter la lecture de l'algorithme, détaillons-en le fonctionnement en supposant que le mot cible est : $x = \text{martin}$. Supposons de plus que la construction de $D(x)$ ait conduit à l'identification de $y = \text{martine}$ comme étant l'élément du lexique partageant la plus longue chaîne initiale avec x (i.e. le premier élément de $D(x)$). On en déduit alors :

$$- \text{pgcp}(x, y) = \text{martin}$$

3. L'idée de contraindre les analogies en leur imposant de s'intégrer à un réseau paradigmatique, en germe dans l'approche mise en œuvre dans le chapitre précédent, est le fruit d'une collaboration avec Vito Pirrelli et Stefano Federici. Qu'ils soient ici remerciés pour leur précieuse contribution, et pour la chaleur de leur accueil aussi. Une partie non négligeable des idées développées dans cette section sont présentées dans (Federici, Pirrelli, & Yvon 1995; 1996). Éric Dauteloup, Marc Lalo, Claire Stricker et Fabien Desserre, étudiants de 2^e année de l'ENST, ont eu la lourde responsabilité de développer une première version du programme mettant en œuvre cette idée. Les résultats que nous présentons ici doivent également beaucoup à la qualité de leur travail.

Algorithme 4: Prononciation par construction de paradigmes

Entrée: Un lexique $\mathcal{L}$, une chaîne $x = c_1 \dots c_{l(x)}$ à transcrire

Sortie: Un ensemble de transcriptions possibles trié de la plus à la moins plausible

PPP(x)

- (1) $\phi(x) \leftarrow \lambda$
- (2) $D(x) \leftarrow \{y \in \mathcal{L} / (t = pgcp(x, y)) \neq \lambda\}$, trié par valeur décroissante de $l(t)$.
- (3) **pour tout** $y \in D(x)$
- (4) $P(x, y) \leftarrow \{(z, t) \in \mathcal{L} \times \mathcal{L} / Pr1(z, t), Pr2(z, t), Pr3(z, t)\}$,
avec :
 - (5) $Pr1(z, t) : z = T_{xy}^F(t)$
 - (6) $Pr2(z, t) : l(pgcp(z, t)) \leq l(pgcp(\phi(z), \phi(t)))$
 - (7) $Pr3(z, t) : l(pgcs(y, z)) \leq l(pgcs(\phi(y), \phi(z)))$
 - (8) trié par valeur décroissante de $l(pgcs(y, z))$
 - (9) **pour tout** $(z, t) \in P(x, y)$
 - (10) **si** $T_{\phi(z)\phi(t)}^F(\phi(y)) \neq \lambda$ **alors**
 - (11) $\phi(x) \leftarrow T_{\phi(z)\phi(t)}^F(\phi(y))$
 - (12) **break**
 - (13) **fin**
 - (14) **fin**
 - (15) **si** $\phi(x) \neq \lambda$ **alors**
 - (16) **break**
 - (17) **fin**
 - (18) **fin**

$$- pgcp(x, y)^{-1}(x) = \lambda$$

$$- pgcs(x, y)^{-1}(y) = e$$

- $T_{xy}^F(a) = b$ définit la transformation orthographique finale applicable à une chaîne finissant comme x de façon à la transformer en une chaîne finissant comme y . Dans le cas présent, cette transformation consiste simplement à rajouter un e à la fin du mot a pour obtenir b . On notera que cette transformation s'applique indépendamment de la forme de a .

L'étape suivante (ligne 4) de l'algorithme consiste à construire P , l'ensemble des couples qui représentent la même alternance que celle exhibée par (x, y) : ici, ce sera la liste de tous les couples (z, t) tels que t se dérive de z par concaténation d'un e final. Ces couples sont ensuite classés suivant la taille de la rime commune à y et t , de manière à examiner en priorité ceux des couples dont le second élément « ressemble » le plus à y . Supposons alors que le mieux classé de ces couples soit :

$$- (z, t) = (\text{marin}, \text{marine})$$

$$- (\phi(z), \phi(t)) = (/mar\tilde{e}/, /marin-/)$$

L'étape finale est de construire la loi de transformation $T_{\phi(t)\phi(z)}^F$ permettant de transformer $\phi(z)$ en $\phi(t)$. On vérifie ensuite qu'elle s'applique à $\phi(y)$: c'est

le sens du *test de compatibilité* de la ligne 10. Dans le cas où la réponse est positive, on peut appliquer cette transformation pour construire $\phi(x)$. Dans le cas présent, cette transformation $T_{\phi(t)\phi(z)}^F$ s'applique à toutes les chaînes phonétiques qui se terminent par /in-/, et remplace ce segment par le segment final /ẽ-/. La transcription du mot *martine* appartient au domaine de définition de cette transformation; en l'y soumettant, on dérive la transcription de *martin* (par $T_{\phi(t)\phi(z)}^F(\phi(y))$): /martẽ-/.

Quelques remarques complémentaires

Condition d'arrêt La condition d'arrêt (ligne 15) telle que nous l'avons décrite est manifestement fâcheuse, dans la mesure où l'exploration des couples de P se termine dès qu'une solution a pu être construite. Ceci fait dépendre la solution de l'ordre dans lequel les paires qui ont une même valeur pour $l(pgcs(y, z))$ sont parcourues; et nous met par là-même à la merci d'une paire ayant un comportement exceptionnel. Nous avons testé trois stratégies alternatives pour contourner ce problème. La première consiste à ne tester l'existence d'une solution qu'à chaque fois que *toutes* les paires de P pour une valeur donnée de $l(pgcs(y, z))$ ont été examinées. Dans ce cas, si plusieurs solutions différentes ont été construites, c'est celle qui est fondée sur le plus grand nombre de paires de P qui l'emporte. La seconde alternative est de remonter le test d'arrêt au niveau du parcours des y , et de produire toutes les solutions pour une taille donnée de $pgcp(x, y)$. Les solutions sont ensuite ordonnées comme précédemment. La troisième stratégie est encore plus onéreuse, puisqu'elle consiste à parcourir complètement D avant de prendre la décision d'arrêt.

C'est la seconde stratégie qui s'est, dans l'ensemble, révélée la plus profitable, et, sauf mention du contraire, les résultats donnés dans les sections suivantes font référence à l'application de cette condition d'arrêt-là.

Auto-sélection de la condition d'arrêt Notons finalement qu'une dernière possibilité aurait été envisageable, qui se fonde sur le raisonnement suivant. Étant donnée une paire (x, y) , et $\mathcal{U}_{x,y}$ l'ensemble des préfixes communs à x et y , chaque u de $\mathcal{U}_{x,y}$ détermine une segmentation de x en uv et de y en uv' . Dans la mesure où l'on cherche à extraire des relations paradigmatiques, on doit s'attendre à ce qu'elles correspondent à une valeur de u telle qu'il existe un grand nombre de paires du type $(u'v, u'v')$. L'idée est alors de surveiller l'accroissement relatif du nombre de solutions produites pour des valeurs croissantes de $l(u)$, et de s'arrêter dès que cette valeur se met à décroître.

Précisons cette intuition sur un exemple précis. Soit le mot cible *morpionner*, néologisme potentiel rendu nécessaire par le succès croissant de la Française des Jeux. Si *morpion* est disponible dans le lexique, il va servir de base à la construction analogique (condition du plus long préfixe). L'algorithme de recherche va alors successivement tester l'existence de segments x telles que les paires suivantes existent dans le lexique :

- x -*morpion* / x -*morpionner* ($l(u) = 0$)
- x -*orpion* / x -*orpionner* ($l(u) = 1$)

- *x-rpion* / *x-rpionner* ($l(u) = 2$)
- *x-pion* / *x-pionner* ($l(u) = 3$)
- *x-ion* / *x-ionner* ($l(u) = 4$)
- *x-on* / *x-onner* ($l(u) = 5$)
- *x-n* / *x-nner* ($l(u) = 6$)
- *x* / *x-ner* ($l(u) = 7$)

Les décomptes effectués sur le lexique BDLEX montrent que la première solution est construite par analogie avec *espion* / *espionner*, qui est la seule paire telle que $l(u)$ vaille 3. Pour $l(u) = 4$, 48 solutions supplémentaires sont produites, sur le patron *X-ion* / *X-ionner*. En augmentant la taille de u d'une nouvelle unité, on obtient 112 nouvelles paires analogues. Le patron *X-n* / *X-ner* ne permettant pas d'améliorer ce score, on peut en déduire que la « bonne » alternance est celle qui correspond au schéma *X-on* / *X-onner*. On notera que, toujours dans ce cas trivial, toutes sauf une (*klaxon*/*klaxonner*) des paires analogues trouvées pour ce niveau de segmentation exhibent la même alternance, qui permet de construire la transcription la plus probable pour *morpionner* : /mɔʁpjɔne/.

Toutefois, dans la mesure où la condition d'arrêt que nous avons imposée induit dans la majorité des cas observés un comportement de l'algorithme identique à ce que fournirait la stratégie d'auto-sélection, nous avons considéré que l'implantation de ce mécanisme, en dépit des principes satisfaisants qu'il met en œuvre, ne permettrait pas d'améliorer de manière suffisamment sensible nos résultats. Il n'a donc pas été implanté.

Ordre de construction Dans notre présentation initiale de l'algorithme, nous avons supposé que le premier terme de la relation analogique était choisi sur la base de la longueur du préfixe partagé avec le mot inconnu. Des expériences complémentaires nous ont convaincu qu'une stratégie plus payante en termes de résultats était d'examiner en priorité les chaînes qui avaient le plus long *suffixe* commun avec la chaîne cible. En conséquence, les résultats que nous listons dans la suite du texte correspondent à une version de l'algorithme pour laquelle tous les *préfixes* de la version de base doivent être remplacés par des *suffixes*, et *vice versa*.

4.3.2 Sur l'utilité de l'alignement

Rime apparente et rime utile

Nous n'avons pas encore élucidé la signification des conditions *Pr2* et *Pr3* des lignes 6 et 7 (voir page 152). Ces conditions ne sont utiles que si l'on dispose de transcriptions *alignées*; auquel cas les symboles « nuls » sont comptés au même titre que les autres dans les calculs de longueur. Ces deux conditions s'interprètent alors comme des contraintes stipulant l'interprétation à donner à ressemblance apparente des chaînes graphiques.

Reprenons l'exemple précédent, et considérons la paire $(z, t) = (\text{romain}, \text{romaine})$. Le second terme partage un suffixe graphique de longueur 3 avec *martine*, et semble donc à mettre sur le même plan que *cousine*. La condition

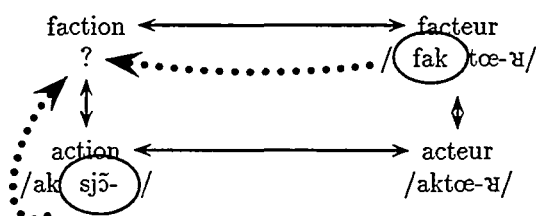


FIG. 4.4 – La segmentation associée à une construction analogique.

Pr3, qui impose que *y* et *t* (ici *martine* et *romaine*) aient un suffixe phonétique de longueur 3 au moins, permet de rejeter cette interprétation, et conduit à considérer que ces deux chaînes n'ont en commun qu'un suffixe de longueur 2 : c'est celui-là qu'il est pertinent d'utiliser dans la construction du paradigme. En d'autres termes, la « bonne » loi de passage entre ces deux mots n'est pas : « remplacer le segment initial *roma* par le segment initial *mart* » ; mais plutôt : « remplacer le segment initial *romai* par le segment initial *marti* ». Par ce biais, l'analogue *romaine* ne sera éventuellement considéré qu'après qu'on aura épuisé toutes les chaînes qui ont vraiment un suffixe commun de taille 3 avec *martine*. Dans les cas favorables, cette paire ne sera alors pas considérée du tout, ce qui est heureux, puisque de toute façon elle ne permet pas de construire une solution (échec au test de compatibilité).

Cet exemple permet de mettre en lumière un aspect intéressant des constructions mises en œuvre par cet algorithme. La donnée d'une paire de mots et des transcriptions afférentes permet d'inférer une segmentation pour chacun des mots, en comparant ce qui, dans les deux chaînes, varie ou reste stable. L'exemple précédent a ainsi montré que la segmentation correcte de *martine*, vue sous l'angle d'un analogue de *romaine*, était *marti+ne* ; et non pas *mart+ine*. La condition *Pr2* s'éclaire alors : elle permet de contrôler dans la construction finale la part qui revient à *t*, et, en conséquence, celle qui doit être prise en charge par *y*, puisque seuls ces deux mots contribuent à la construction de $\phi(x)$. Considérons l'exemple suivant, reproduit sur la figure 4.4.

Dans ce nouvel exemple, le mot cible est $x=faction$, que l'on cherche à transcrire par analogie avec le mot $y=facteur$; le couple de mots qui sert de support à l'analogie est la paire $(z,t)=(action,acteur)$. Ces deux derniers mots ont en commun un préfixe de longueur 3 : *act*. Néanmoins, dans la mesure où ce segment est transcrit différemment dans les chaînes phonétiques correspondantes, la condition *Pr2* impose de considérer que le plus long préfixe commun à ces deux mots est en fait de taille 2. Il en découle une segmentation pour la chaîne résultante : /faksj̃-/ se décompose en un segment /fak/, inspiré de la transcription de *y*, et un segment /sj̃-/ , inspiré de la prononciation de *z*. Similairement, *faction* se segmente en *fac+tion*.

En résumé, ces deux conditions permettent, lorsque des alignements sont disponibles, d'ordonner plus finement les ensembles *D* et *P*. La conséquence la plus manifeste de ce réordonnement est que l'examen des solutions les moins prometteuses est différé (ce qui, dans la version de base de l'algorithme, équivalait à un filtrage, puisque l'on stoppe dès la première solution trouvée).

Transcrire sans aligner

Il est intéressant dans ce contexte de s'interroger sur l'utilité de posséder des transcriptions alignées pour développer cette stratégie. Dans la mesure où toutes les opérations sur les chaînes portent sur des segments initiaux ou terminaux, la disposition d'un alignement est en fait un luxe. Pour nous en convaincre, considérons à nouveau l'exemple de la section 4.3.1. La recherche de y comme celle de la paire (z, t) se fonde uniquement sur l'aspect des chaînes graphiques, et ne présuppose en aucune manière l'existence d'un alignement. Le calcul de $T_{\phi(z)\phi(t)}^F$ ne dépend pas non plus de cette information : si on l'omet, cette transformation se glose en « remplacer un segment /in/ final en /ē/ » (à comparer avec l'interprétation donnée en page 153). Par rapport à la version avec alignement, seul le domaine de définition de la transformation a changé : elle s'applique maintenant à un mot comme *amin-/amin/* qui en était auparavant exclu. Le résultat, par contre, ne change pas, puisqu'en l'appliquant à la prononciation de *martine*, on obtient la prononciation désirée.

Il est toutefois certain que la connaissance des alignements permet une double amélioration des performances. Une amélioration de la correction, puisque l'alignement permet de préciser le domaine de définition des transformations, et, par là même, de contraindre le nombre de solutions construites. Amélioration également en ce qui concerne le temps de traitement, puisque, nous l'avons vu, utiliser des alignements autorise à tenir compte de la taille utile des segments partagés, et non pas simplement de leur taille apparente.

On notera toutefois que nous nous éloignons singulièrement sur ce point de la démarche classificatoire : l'alignement n'est plus une condition préalable mais simplement un moyen d'améliorer à la marge les résultats. En étendant la démarche présentée ici (voir la section 4.4), nous nous affranchirons complètement de cette contrainte, et ne considérerons plus que des transcriptions non alignées.

4.3.3 Résultats

Évaluation sur différents lexiques

Les performances de PPP sur différents lexiques détaillées dans la table 4.2 témoignent du succès de la stratégie mise en œuvre par cet algorithme vis-à-vis des objectifs que nous nous étions fixés à la section 3.2.5.

En effet, à l'examen de ces résultats, on constate en premier lieu que nous avons réussi à augmenter très sensiblement, par rapport aux approches précédemment rapportées, le pourcentage des mots correctement transcrits (la justesse); et que cette amélioration est à peu près uniforme sur tous les lexiques. Corollairement, on observe toutefois une très nette augmentation du silence, qui, dans le cas du lexique CMU-20 atteint des proportions inacceptables, puisque près de deux mots sur trois ne sont tout simplement pas prononcés. Le cas du lexique TUB est à cet égard encore pire, puisque que seuls un nombre marginal de mot sont transcrits.

Soulignons, et ces résultats en attestent, que la recherche de paradigmes telle qu'elle est ici exemplifiée diffère profondément des algorithmes décrits dans les

Corpus	Par mot			Par symbole	
	Correction	Silence	Justesse	Correction	Justesse
NETTALK	50,460	31,420	73,586	61,840	95,197
NETSTRESS	49,305	31,330	71,791	61,020	94,763
BDLEX	59,115	35,953	92,297	60,200	98,770
CMU-20	24,494	63,848	67,746	28,994	93,358
ENST-20	70,850	19,312	87,800	77,252	97,332
ILC-20	59,045	38,693	96,312	57,130	99,442
TUB	1,879	97,682	81,144	1,698	96,086
KUN	68,338	27,592	94,375	68,656	99,095

TAB. 4.2 – Performances de PPP sur différents lexiques.

chapitres précédents. Les performances dépendent maintenant crucialement de la nature du lexique d'exemples qui est appris, dans la mesure où le succès d'une construction est soumis à la présence dans le lexique d'apprentissage d'un mot voisin (i.e. qui forme avec lui une pseudo-paire minimale), ou d'un composé morphologique de même racine. La stratégie de construction « au hasard » qui a prévalu pour bâtir les lexiques CMU-20, KUN, TUB et ILC-20 est dans ce contexte désastreuse, puisqu'elle a consisté à sélectionner aléatoirement des entrées dans une base de large couverture (la couverture la plus large étant sans aucun doute celle du lexique TUB, dont les entrées ont été choisies au hasard dans un lexique de près d'un million de noms propres, et à la lecture duquel il est difficile de deviner que l'on a affaire à un lexique de noms allemands). Les taux de silence plus faibles observés sur le lexique ENST-20, et, dans une moindre mesure, sur les lexiques NETTALK et BDLEX témoignent au contraire qu'il est préférable de travailler sur un lexique construit sur des bases fréquentielles. Par ce biais, on est assuré que les racines principales, à partir desquelles de nombreux termes moins fréquents sont dérivés, sont effectivement disponibles pour la transcription.

Comparaison avec les approches alternatives

En comparant PPP avec les autres algorithmes dont nous disposons, il s'avère que l'approche analogique conduit toujours à des résultats globaux sensiblement inférieurs aux autres approches, conséquence directe d'un taux de silence par trop important. Ce qui, par contre, est encourageant, c'est que le mécanisme de constriction paradigmatique se révèle particulièrement adapté à la détection des mots qui sont susceptibles de poser problème : ainsi, nos expériences avec NETTALK ont mis en évidence le fait que près de la moitié (46,42 %) des mots incorrectement transcrits par DEC (le même vaut pour SMPA) ne sont pas prononcés par PPP.

Réciproquement, les matrices de confusion entre les différents algorithmes permettent de chiffrer les gains d'une stratégie mixte utilisant plusieurs algorithmes : utiliser PPP pour transcrire peu de mots, mais avec une grande correction ; et utiliser un des autres algorithmes pour suppléer aux manques de la stratégie paradigmatique (à la manière du fonctionnement de TTs (Coker, Church, & Liberman 1990), où un système à base de règle est utilisé dans les cas où l'approche analogique échoue). Dans le cas de SMPA comme dans le cas

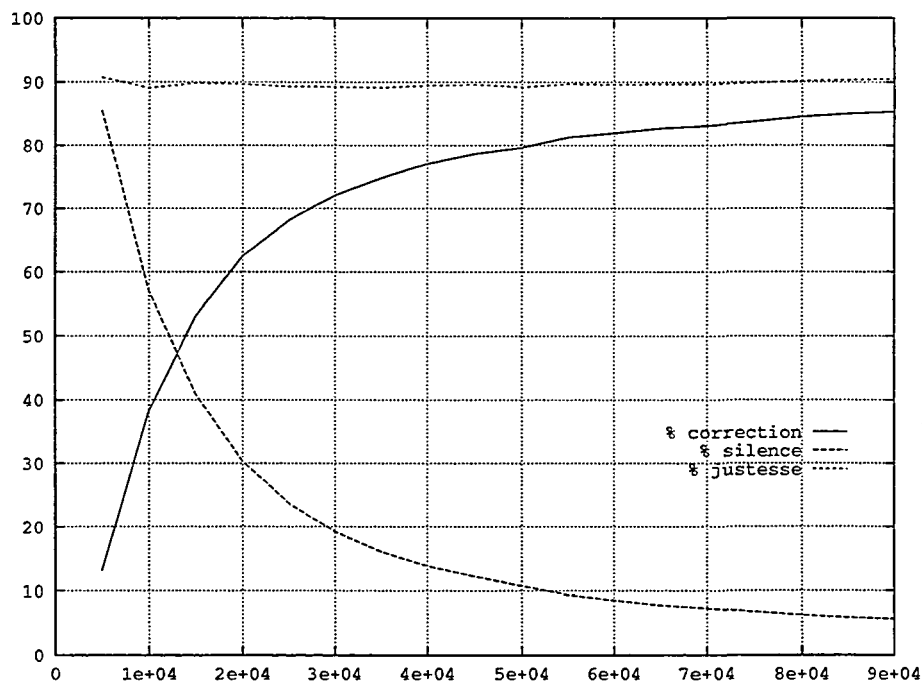


FIG. 4.5 – Évaluation de la vitesse d'apprentissage de PPP sur le lexique ENST.

de DEC ce mode de coopération conduit à des résultats légèrement meilleurs que l'utilisation d'un seul algorithme, conduisant dans les deux cas à des taux de correction par mot de l'ordre de 65 %.

Comportement à la limite

Une autre caractéristique réellement intéressante de PPP est reflétée par l'examen des courbes d'apprentissage sur des lexiques volumineux, telle que celle présentée à la figure 4.5.

Ce que ce graphique met en évidence, c'est l'extrême prudence de la stratégie paradigmatique telle que nous l'avons implantée. Si le lexique d'exemples est petit, le nombre de mots transcrits sera très faible, mais si on fournit un nombre d'exemples suffisant, alors le taux de silence diminue pour retrouver des proportions beaucoup plus raisonnables (il tombe en dessous de 10 % dès que le lexique d'apprentissage excède 50 000 noms). Par contre, le niveau de justesse des prédictions est pratiquement indépendant de la taille du lexique d'apprentissage : qu'il soit de 5 000 ou de 95 000 mots, environ 9 noms sur 10 pour lesquels une transcription est produite sont corrects. En d'autres termes, PPP est tout le contraire d'omniphone : il prononce avec une circonspection fortement dépendante du volume de connaissance disponible, mais le fait toujours fort judicieusement. La comparaison avec la courbe d'apprentissage de DEC sur le même lexique témoigne également du fait que PPP ne cesse pas de tirer profit des exemples supplémentaires, ce que DEC ne fait pas (c.f. le graphique B.2 de

la page 233).

Des conclusions identiques se dégagent de l'examen de la courbe d'apprentissage, non reproduite ici, obtenue en utilisant le lexique CMU. Dans ce dernier cas, l'avantage quantitatif de la stratégie paradigmatique pour traiter de gros volumes de données apparaît avec plus de relief : les performances de PPP lorsque 100 000 mots sont appris sont de 56,73 % de correction et 66,15 % de justesse, avec un taux résiduel de silence inférieur à 15 %. À titre de comparaison, pour une taille équivalente du lexique d'apprentissage, le taux de correction par mot de DEC est de l'ordre de 46,38 %, soit plus de 10 points inférieur.

Filtrage morpho-syntaxique des paradigmes

Principe Le lexique BDLEX présente la particularité, par rapport aux autres lexiques dont nous disposons, de spécifier, pour chaque entrée, la catégorie syntaxique majeure (voir l'annexe F). Cette spécification est relativement frustrante, puisqu'elle distingue seulement cinq catégories pour les mots simples : **Verbe**, **Nom**, **Adjectif**, **Adverbe**, et **Autre**. Cette information supplémentaire nous a permis de mesurer l'impact sur les performances de notre algorithme du raffinement⁴ qui consiste à poser une condition de compatibilité morpho-syntaxique sur les entrées lexicales considérées comme analogues et utilisées pour construire la prononciation de termes inconnus.

Précisons graphiquement, sur un diagramme de commutation, la manière dont s'exerce cette contrainte. Pour un mot cible x , PPP recherche une segmentation de x en deux morceaux $x = au$, et deux segments auxiliaires b et v tels que les mots bu , av et bv appartiennent au lexique. La condition d'agrément morpho-syntaxique surimpose la contrainte que au et bu d'une part, av et bv d'autre part, s'accordent sur leur catégorie, comme sur le graphe de commutation contraint représenté à la figure 4.6, où l'on a noté $c(x)$ la catégorie morpho-syntaxique associée à l'entrée x .

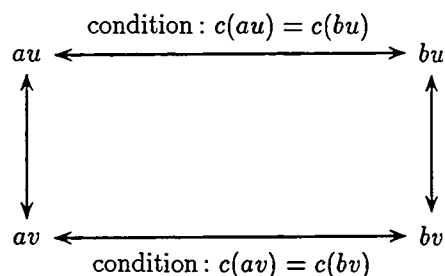


FIG. 4.6 – Un graphe de commutation morpho-syntaxiquement contraint.

Une autre manière de poser cette contrainte aurait été d'exiger que l'accord portât d'une part sur les catégories de au et av d'une part, et de bu et bv d'autre part. Dans la mesure où toutefois cette contrainte était spécifiquement conçue pour le traitement d'un lexique français, notre choix se justifie assez naturellement. En français, les dérivations morphologiques s'expriment

4. Ce raffinement est également indépendamment suggéré, mais non-évalué, dans (Pirrelli & Federici 1995).

en effet préférentiellement au travers de mécanismes de suffixation; ce sont en conséquence les segments terminaux des mots qui sont généralement les plus indicatifs de la catégorie morpho-syntaxique. Donnons un exemple du fonctionnement de cette contrainte en considérant le mot cible *cancer*. En l'absence de contrainte, ce nom a toutes les chances d'être prononcé par analogie avec des verbes tels que *lancer* ou *tancer*. La contrainte que le terme de même suffixe dans le schéma paradigmatique soit de même catégorie morpho-syntaxique va éliminer ces constructions (qui mènent à des solutions incorrectes), et ainsi favoriser des constructions élaborées, pour leur partie finale, à partir d'autres noms tels que *canter* ou *pater*.

Notons que ce type de contrainte peut être étendu en tenant compte non seulement des catégories, mais également des *traits morpho-syntaxiques*, pour l'adapter à des lexiques comportant des formes fléchies. Pour reprendre un exemple maintenant familier, celui de *Martin*, on note qu'un tel mécanisme pourrait être formalisé en imposant que le prénom partageant un suffixe commun avec *Martin* soit un prénom *masculin*; et que de même le prénom partageant un suffixe commun avec *Martine* soit un prénom *féminin*. Dans le même ordre d'idée, l'utilisation de telles contraintes est envisageable pour transcrire des noms propres qui seraient étiquetés par une indication de leur origine linguistique: ainsi un nom identifié comme étant d'origine française ne pourrait avoir comme analogue pour sa partie finale qu'un autre nom de même origine.

Résultats Les résultats sont relativement conformes à ce que l'on était en droit d'espérer: la constrictio des analogues permet une amélioration sensible des résultats, puisque la correction par mot augmente de plus d'un point (60,43% contre 59,11%), et la justesse d'environ 2 points, s'établissant à un niveau voisin de 95%. De plus, dans le cas de BDLEX, ce gain de performances s'effectue sans augmentation du silence, témoignant de la « densité paradigmatique » de ce lexique: filtrer un certain nombre de couples d'analogues n'épuise pratiquement jamais l'ensemble des candidats possibles.

4.3.4 Discussion et perspectives

Nous avons présenté un algorithme de prononciation par analogie, fondé sur un principe novateur et linguistiquement fondé, celui de l'extraction de paradigme. Cet algorithme nous a permis d'unifier sur des bases solides deux visions de l'analogie *a priori* divergentes: celle mise en avant dans les modèles cognitifs de la lecture, correspondant à une ressemblance mal spécifiée, et celle plus formalisée qui prédomine en intelligence artificielle; et en particulier de donner un sens précis à la notion d'analogue (dans le cadre de la tâche de lecture).

Les résultats que nous avons obtenus en appliquant cette approche analogique sur une batterie de lexiques ont mis en évidence deux caractéristiques principales qui l'opposent aux techniques présentées dans les chapitres précédents: d'une part un taux de non-prononciation très important, d'autre part un taux de justesse sur les mots transcrits bien supérieur à ce que nous avons pu obtenir par ailleurs.

Cette double caractéristique suggère deux directions de recherche pour amé-

4.4 Prononcer par application de transductions paradigmatiques 161

liorer les performances de l'approche paradigmo-analogique: d'une part essayer d'augmenter à la marge les scores de correction; d'autre part essayer de diminuer les taux de silence, tout en conservant le principe général d'analogie. C'est sur cette dernière voie de réflexion que nous entraîne la section suivante; choix motivé d'une part par le besoin de produire des mécanismes de transcription réellement opératoires, y compris sur des lexique de taille moyenne; et d'autre part par le sentiment que la mise en œuvre dans PPP du principe paradigmatique est trop simpliste pour capturer de façon satisfaisante l'ensemble des régularités qui structurent les lexiques de formes graphiques. Avant de nous lancer, suggérons toutefois quelques idées qui nous semblent dignes d'approfondissement pour améliorer la justesse de nos transcriptions.

Nous l'avons déjà étudié, l'intégration, même dans un modèle très frustré, de contraintes morpho-syntaxiques est un moyen effectif de parvenir à ce but. On peut également envisager de relaxer les contraintes de stricte concordance graphique qui pour l'instant déterminent exclusivement les analogies, en introduisant par exemple des classes de lettres *a priori* analogues. C'est la voie suivie par Pirrelli et Federici (1995), qui mentionnent une amélioration corrélative de l'ordre de 0,5 % des performances. Une troisième direction de recherche consiste à fonder l'approche analogique non plus sur les chaînes graphiques, mais sur les chaînes phonétiques: un moyen simple de mettre ce principe en œuvre serait d'utiliser PPP de conserve avec un autre système de transcription capable de produire pour chaque entrée plusieurs solutions; dans cette optique PPP serait déchargé de la tâche de transcription, et ne serait plus en charge que de leur reclassement, en choisissant celle qui présente le meilleur ancrage dans un système de relations paradigmatiques. Cette manière d'exploiter les relations analogiques comme un mécanisme de classement de transcription concurrentes se rapproche de celle mise en œuvre par Sullivan et Damper (1990; 1991; 1992), qui utilisent toutefois un principe de reclassement différent de celui suggéré ici, conforme avec leur conception de l'analogie comme un principe de ressemblance topologique.

4.4 Prononcer par application de transductions paradigmatiques

Dans cette section, nous présentons une extension de l'approche paradigmatique visant à assouplir les conditions de prononçabilité d'un mot. Nous exposons dans un premier temps les motivations linguistiques qui justifient cette approche, et introduisons l'idée générale de la démarche, fondée sur un principe de cheminement paradigmatique. Nous présentons ensuite plus formellement l'algorithme implantant cette idée. Des résultats préliminaires sont ensuite fournis, qui permettent principalement de montrer, que, malgré des défauts certains de mise en œuvre, cette approche est susceptible de fournir des performances de transcription suffisamment satisfaisantes pour que des extensions variées soient envisagées avec optimisme.

4.4.1 Couplage de transformations paradigmatiques

Une reformulation de ppp

La stratégie mise en œuvre dans PPP peut se reformuler différemment, en la considérant sous l'angle des transformations lexicales. Reprenons à cet effet l'exemple de *faction* discuté à la section 4.3.2. L'extraction du schéma analogico-paradigmatique conduisant à la transcription de ce mot est indissociable de l'identification du couplage deux à deux de quatre paires de transformations, chaque paire de transformations contenant une transformation dans le domaine graphique, et une transformation dans le domaine phonétique :

1. $f \rightarrow \lambda$; $/f/ \rightarrow \lambda$;
2. $\lambda \rightarrow f$; $\lambda \rightarrow /f/$;
3. $ion \rightarrow eur$; $/sj\tilde{5}/ \rightarrow /t\text{œ}\mathfrak{A}/$;
4. $eur \rightarrow ion$; $/t\text{œ}\mathfrak{A}/ \rightarrow /sj\tilde{5}/$;

On notera, de plus, que ces transformations sont deux à deux (1 et 2, 3 et 4) en relation d'inverse, et que chaque diagramme de commutation correspond à l'identification, pour chacune des transformations, de deux éléments du lexique dont le transformé est également dans le lexique : ainsi la première transformation est attestée par les couples (*action*, *faction*) et (*acteur*, *facteur*).

On notera également que l'on peut reconnaître au fondement de ces transformations deux principes très différents. Les deux premières, en première approche, peuvent être justifiées par un principe de densité lexicale des radicaux. Dans la mesure où les racines sont en nombre important, et de longueur généralement faible, la forte densité lexicale qui résulte permet de prédire que de nombreuses transformations de ce type (substitution d'une chaîne initiale) seront observées. C'est ainsi le cas pour pratiquement toutes les transformations en français qui échangent deux groupes de consonnes (éventuellement vides) initiaux. Nous ferons référence dans la suite de ce document à ces transformations sous le nom de transformations *radicales*⁵.

Les deux secondes, elles, peuvent être vues comme découlant de la productivité lexicale de la langue, qui permet qu'un grand nombre d'uffixes se greffent sur les racines disponibles suivant des schémas plus ou moins contraints (phonologiquement, syntaxiquement ou sémantiquement). Nous référerons à ces dernières par le terme de transformations *affixales*.

Reformulé dans ces termes, PPP est capable de prononcer un mot x inconnu si et seulement le lexique d'apprentissage contient :

- une forme y dérivable par transformation affixale T_a .

5. L'utilisation de ce terme est largement abusive. Notre analyse d'une forme ne se fonde que sur des ressemblances/dissembances purement graphiques; elle est à ce titre incapable d'identifier de manière précise des unités telles que radicaux ou des affixes. Nous conserverons néanmoins ce terme (comme celui d'affixal introduit plus loin), qui, nous semble-t-il, permettent de saisir l'intuition de notre démarche.

4.4 Prononcer par application de transductions paradigmatiques 163

- un couple (z, t) tel que z se dérive de t par T_a ; et il existe une transformation radicale T_l qui permet de construire z à partir de x , et t à partir de y .

Cet ensemble de conditions est relativement contraignant, comme en témoignent les taux de silence obtenus en expérimentant PPP sur les différents lexiques. Il semble donc difficile de se reposer sur ce seul mécanisme pour fonder, avec des taux de succès raisonnables, un système de transcription.

À propos des « trous » dérivationnels, et comment les boucher

La nécessité de complexifier le mécanisme mis en œuvre dans PPP, afin en premier lieu de réduire les taux de silence, résulte de la présence de trous dans les données dont nous disposons. Ces trous, dans une large proportion, résultent de la taille finie de nos lexiques. Ainsi BDLEX connaît *poison* et *empoisonner*, mais pas *poissonner*, qui joue ainsi le rôle de « chaînon-manquant » entre ces deux termes. De cette absence, résulte l'impossibilité pour PPP de prononcer aussi bien l'un que l'autre des deux mots sur des bases correctes. Ceci est naturellement encore plus flagrant pour les mots importés, qui abandonnent bien souvent, dans l'expatriation, le voisinage lexical familial et familial au sein duquel ils prennent tout leur sens.

À ce facteur contingent, s'ajoute un second facteur linguistique qui motive de manière plus radicale le raffinement de nos procédures. Dans les lexiques (de formes non-fléchies) que nous traitons, les transformations affixales les plus productives sont en effet interprétables en termes de dérivations morphologiques. La morphologie dérivationnelle se distingue toutefois, entre autres traits, de la morphologie flexionnelle, par l'extrême porosité des paradigmes. Nombreuses sont ainsi les transformations lexicales interprétables en termes de dérivation qui s'appliquent de façon restrictive à un petit sous-ensemble des formes possibles, formant pour ainsi dire des *trous*⁶ dans le paradigme.

Un certain nombre de motifs peuvent être invoqués pour justifier l'absence dans le lexique de formes que l'on s'attendrait à y trouver, compte-tenu de l'existence de schémas de construction réguliers.

Dans de nombreux cas, tout d'abord, l'existence de formes supplétives rend la formation du dérivé inutile. Ainsi, à l'inverse de *curious* duquel on dérive *curiosity*, on ne peut dériver **gloriosity* de *glorious*, parce que *glory* est déjà disponible dans le lexique.

Une seconde justification de ces trous est l'existence de *conditions de sélection* attachées aux dérivations affixales. Ainsi en anglais, si certains affixes peuvent modifier n'importe quel élément d'une catégorie morpho-syntaxique donnée (ainsi *-ness*); d'autres (*-ify*, *-hood*) ne peuvent s'attacher qu'à un radical qui n'a pas été déjà affixé. Une troisième classe de suffixes peut s'attacher

6. Le terme de *trous* s'applique quelque peu improprement aux phénomènes dérivationnels que nous décrivons ici. Les trous n'existent réellement que dans les paradigmes inflexionnels, dans lesquels ils sont (quasi) systématiquement comblés par des formes supplétives. Dans le cas des dérivations, parler de trous est impropre, puisque certaines constructions sont interdites. Ce sont, pour ainsi dire, des trous *naturels*.

librement à une racine, ou à certaines formes précédemment dérivées : une condition de compatibilité entre les deux affixes limite néanmoins les constructions possibles; ainsi *-ic* ne peut modifier un terme dérivé que s'il est de la forme *-ist* pour former des composés en *-ist-ic*.

Des conditions de sélection phonologiques limitent la productivité de certains paradigmes dérivationnels. Ainsi, toujours dans le cas de l'anglais, la formation de verbes à partir d'adjectifs par suffixation de *-en* (sur le schéma :*red-redden*) n'est possible que pour des adjectifs monosyllabiques qui se terminent par une obstruante. D'autres conditions de sélection trouvent leur origine dans l'histoire de la langue; ainsi la nominalisation par suffixation de *(a)tion* en anglais n'est autorisée que pour des verbes d'origine latine.

Il existe enfin des trous semble-t-il inexplicables, des mots dont on ne peut que constater l'absence (du moins exprimer de sérieux doutes sur leur acceptabilité) : ainsi des formes comme *charmamment* (dérivé de *charmant*), ou encore *clémement* (dérivé de *clément*) semblent douteuses, malgré l'existence des paires (*élégant, élégamment*) et (*décent, décemment*) (Molinier 1992).

L'existence de trous lexicaux, aggravée par les phénomènes d'allomorphie (le principe de similitude graphique sur lequel nous nous fondons nous interdit ainsi de détecter des alternances telles que *remit/commit, reduce/conduce*) limite *théoriquement* les potentialités d'une approche telle que PPP, quand bien même on la nourrirait de lexiques gigantesques. Prenant acte de cette limitation, nous nous sommes alors attachés à construire un mécanisme qui permette, tout en conservant un rôle central au concept de paradigme, garant de taux de justesse importants, d'assouplir les conditions de prononçabilité d'une forme inconnue. En d'autres termes, nous nous sommes attachés à trouver un mécanisme qui permette de dériver *empoisonner* de *poison*, y compris en l'absence de *poisonner*.

Boucher les trous dans une dérivation complexe

La réponse que nous avons apporté à ce problème s'exprime de manière très simple: bouchons les trous. Mettre en œuvre ce mot d'ordre sur l'exemple de *empoisonner*, nécessite tout d'abord d'être capable de réaliser que *poisonner* (ou d'ailleurs *empoison*) est un terme utile, à partir duquel *empoisonner* se dérivera simplement par une préfixation bien attestée; d'être ensuite capable de simuler la présence de cet élément dans le lexique, en en produisant notamment une prononciation. Nous disposons pour cette deuxième étape, avec PPP, d'un mécanisme qui a fait ses preuves. Cette démarche est schématisée dans le graphique suivant (figure 4.7), qui présente le double graphique de commutation permettant d'inférer de *poison* la prononciation de *poisonner*.

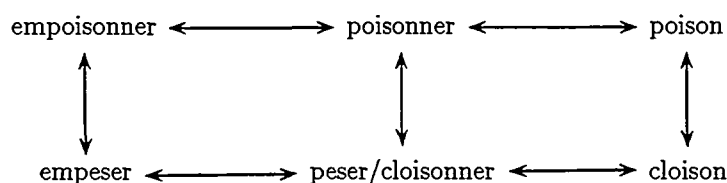


FIG. 4.7 – Un double graphe de commutation.

4.4 Prononcer par application de transductions paradigmatiques 165

Toute la difficulté de cette approche réside clairement dans l'identification correcte et rapide des « chaînons manquants » (dans notre exemple *poisonner*, mais *empoison* ferait tout aussi bien l'affaire), qui doivent simultanément entrer dans un système d'alternance paradigmatique avec le mot inconnu et avec au moins un mot du lexique.

Un certain nombre d'approches différentes à ce problème étaient à ce stade disponibles : nous avons, par exemple, envisagé de commencer par choisir sur des critères de recouvrement maximal le mot du lexique le plus proche du mot inconnu, pour disposer *a priori* de la forme graphique des éléments manquants. Dans l'exemple, cela conduirait à identifier immédiatement *poison*, puis à en déduire les transformations à opérer pour construire la prononciation inconnue.

Notre choix s'est finalement porté sur une approche permettant de ne pas limiter *a priori* le nombre de chaînons entre le mot inconnu et le mot du lexique. Pour en fixer immédiatement l'intuition, disons que l'on cherchera à construire tous les dérivés paradigmatiques possibles du mot inconnu. Si aucun d'entre eux n'est dans le lexique, alors on s'intéressera aux dérivés des dérivés, puis aux dérivés des dérivés des dérivés, etc. La section suivante est consacrée à la formalisation de cette démarche.

4.4.2 Prononcer par cheminement paradigmatique

L'idée générale dont nous allons maintenant expliciter la mise en œuvre est la suivante : lorsqu'un mot de prononciation inconnue est proposé au système, il lui est appliqué « à l'aveugle » des transformations graphiques jusqu'à ce que l'on reconstruise un mot connu (dans la suite appelé *analogue*). Une fois l'analogue trouvé, on applique à sa transcription une série de transformations dans le domaine phonétique, qui reproduit à l'inverse le parcours transformationnel effectué dans le domaine graphique. Nous avons baptisé cette démarche *cheminement paradigmatique*.

Appliquée à *empoisonner*, cette démarche idéalement produira *poisonner* (ou *empoison*), duquel sera dérivé *poison*, l'analogue. Il s'agira ensuite d'appliquer à /pwaz5/ des transformations phonétiques « inverses » (dans un sens qui reste à préciser), pour obtenir la prononciation souhaitée.

Dans cette section, nous présentons formellement la procédure effectivement mise en œuvre pour implanter la démarche de *cheminement paradigmatique*. Cette procédure requiert que l'on soit capable d'identifier un jeu de transformations pertinent (d'identifier sur notre exemple, que le mécanisme de préfixation/dépréfixation de *em* est productif); que l'on soit capable de se munir d'un ordre d'évaluation des transformations (et des transformés); et finalement que l'on soit capable de donner un sens précis au concept de *transformation phonétique inverse*. C'est le propos des paragraphes qui vont suivre, dans lesquels, après avoir introduit quelques notations, nous détaillons, étape par étape, le fonctionnement de notre algorithme, en motivant et illustrant les différents choix d'implantation que nous avons été conduit à faire.

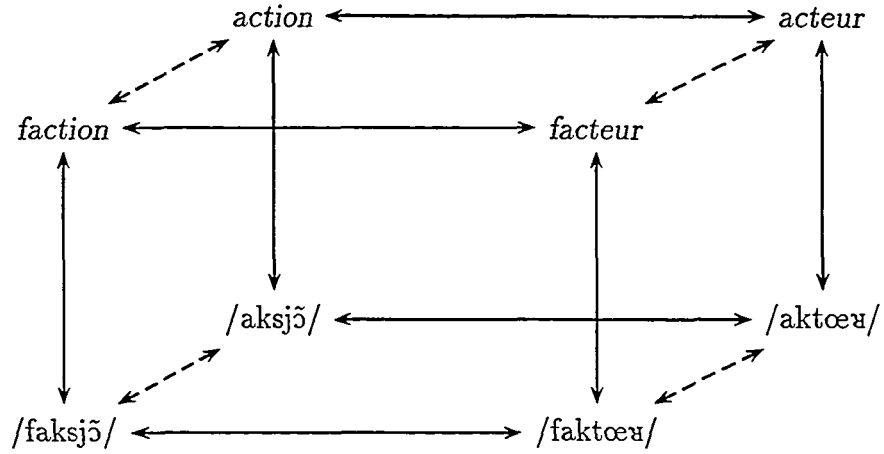


Figure 4.8: Un graphe de commutation tri-dimensionnel.

Introduction formelle

Un certain nombre de remarques formelles⁷ sont ici nécessaires, ainsi que l'introduction de nouveaux termes, qui nous permettront d'exposer plus précisément dans les prochains paragraphes la manière dont nous avons mis en œuvre algorithmiquement l'idée de cheminement paradigmatique.

L'exemple qui nous servira de référence au cours de cette introduction formelle est celui, déjà utilisé à la section 4.3.2, du système formé par les lexies *faction*, *facteur*, *action* et *acteur*. Ce système d'alternance est reproduit à la figure 4.8.

Notations Une transformation T_G de $\mathcal{G}$ dans lui-même est définie par un triplet (les paramètres de T_G) (a, b, X) de $\mathcal{G} \times \mathcal{G} \times \{I, F\}$ de la manière suivante: $\forall(x) \in \mathcal{G}$,

$$\left\{ \begin{array}{ll} \text{si } X = I : & \begin{array}{l} \text{si } \exists u \in \mathcal{G} \text{ tel que } x = au \\ \text{alors } T_G(x) = bu \\ \text{sinon } T(x) \text{ n'est pas défini} \end{array} \\ \text{si } X = F : & \begin{array}{l} \text{si } \exists u \in \mathcal{G} \text{ tel que } x = ua \\ \text{alors } T(x) = ub \\ \text{sinon } T(x) \text{ n'est pas défini} \end{array} \end{array} \right.$$

On appelle X le *type* de la transformation T_G , et on parlera de transformation *initiale* (respectivement *finale*) dans le cas où $X = I$ (resp. $X = F$) et on notera $\mathcal{T}_G^I$ (resp. $\mathcal{T}_G^F$) l'ensemble de transformations correspondant. On définit de même les transformations T_P de $\mathcal{P}$ dans lui-même. On notera finalement D_{T_G} (resp. D_{T_P}) le sous-ensemble de $\mathcal{G}$ (resp. $\mathcal{P}$) sur lequel une transformation T_G (resp. T_P) est définie (son *domaine de définition*).

⁷. Dans ce paragraphe, on « oublie » pour faciliter la formalisation, que ϕ n'est pas, en général, une fonction (c.f. la discussion de la page 11).

4.4 Prononcer par application de transductions paradigmatiques 167

Une (co-)transformation⁸ T de $\mathcal{L}$ dans $T(\mathcal{L})$ est définie comme un couple de transformations (T_G, T_P) de même type : ce sont les transformations associées à T . Une transformation est caractérisée par un quintuplet (a, b, α, β, X) de $\mathcal{G} \times \mathcal{P} \times \mathcal{G} \times \mathcal{P} \times \{I, F\}$ de la manière suivante (on suppose que T_G (resp. T_P) a (a, b, X) (resp. (α, β, X)) pour paramètres) : $\forall (x, \phi(x)) \in \mathcal{L}$,

$$\begin{cases} \text{si } x \in D_{T_G} \text{ et } \phi(x) \in D_{T_P} \text{ alors } T(x, \phi(x)) = (T_G(x), T_P(\phi(x))) \\ \text{sinon } T(x, \phi(x)) \text{ n'est pas définie} \end{cases}$$

De manière naturelle, on notera D_T le domaine de définition d'une transformation, à savoir l'ensemble des paires (x, y) de $\mathcal{G} \times \mathcal{P}$ telles que x soit dans D_{T_G} et y dans D_{T_P} .

Quelques définitions et propriétés Une transformation T est *attestée* sur un lexique $\mathcal{L}$ si et seulement si il existe deux éléments du lexique $(x, \phi(x))$ et $(y, \phi(y))$ tels que $T(x, \phi(x)) = (y, \phi(y))$. Une formulation équivalente, qui a l'avantage de faire réapparaître les diagrammes de commutation (voir figure 4.9), est la suivante : $T = (T_G, T_P)$ est attestée si et seulement si il existe un élément x de $\mathcal{L}$ qui vérifie : $\phi(T_G(x)) = T_P(\phi(x))$. On dira dans ce cas que T *P-commute* avec ϕ sur $\{x\}$.

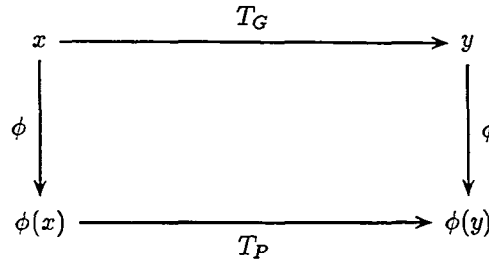


FIG. 4.9 – Un graphe de commutation abstrait.

Une transformation est *paradigmatique* sur $\mathcal{L}$ si et seulement si elle est doublement attestée par les couples (x, y) et (z, t) ; ou encore si le sous-ensemble L_T de $\mathcal{L}$ sur lequel T P-commute avec ϕ est de cardinal supérieur ou égal à 2. On notera $\mathcal{T}_{\mathcal{L}}$ le sous-ensemble de $\mathcal{T}$ paradigmatique sur $\mathcal{L}$.

Par exemple, le système représenté à la figure 4.8 permet d'attester doublement la transformation initiale dont les composantes graphiques (resp. phonétiques) sont : $a \rightarrow fa$ (resp. $/a/ \rightarrow /fa/$).

On notera que :

Propriété 4.4.1 (Trivialité) pour tout (x, y) de $\mathcal{L} \times \mathcal{L}$, la transformation T de paramètres $(x, y, \phi(x), \phi(y), I)$ est attestée, de même que T' la transformation finale de paramètres $(x, y, \phi(x), \phi(y), F)$.

8. Nous parlerons dans la suite de cette introduction de transformation *graphique* et de transformation *phonétique* pour référer aux éléments de respectivement T_G et T_P . Nous réserverons le terme de transformation pour désigner les co-transformations.

On appellera transformation *triviale* une transformation dont l'ensemble de définition sur $\mathcal{L}$ est un singleton. Dans la suite, sauf mention explicite du contraire, il faudra entendre par transformation une transformation non-triviale.

Propriété 4.4.2 (Symétrie) *si T , de paramètres (a, b, α, β, X) , est attestée (resp. paradigmaticque), alors T^{-1} de paramètres (b, a, β, α, X) est également attestée (resp. paradigmaticque).*

Propriété 4.4.3 (Dualité) *si T , de paramètres $(x, y, \phi(x), \phi(y), F)$, est paradigmaticque sur $\mathcal{L}$, alors T' définie par les paramètres $(u, v, v, v, \overline{X})$ (avec : $x = au, y = bu, z = av, t = bv, \phi(x) = \alpha v, \dots$, et $\overline{X} = I$ si $X = F$ et F sinon) est également paradigmaticque.*

Cette propriété de dualité transparaît clairement sur la représentation graphique de la figure 4.8 : le plan de la feuille représente un couple de paires attestant la transformation échangeant un *a* avec un *fa* initial; dans le plan orthogonal à la feuille on trouve un couple de paires attestant la transformation finale échangeant graphiquement les suffixes *ction* et *cteur*, et phonétiquement les suffixes /ksjõ/ et /ktœʁ/.

Propriété 4.4.4 (Indétermination) *si T est attestée pour (x, y) sur $\mathcal{L}$, alors (avec les notations usuelles) $\forall (c, \gamma) \in \mathcal{G} \times \mathcal{P} / \lambda \leq c \leq pgcp(x, y)$ et $\lambda \leq \gamma \leq pgcp(\phi(x), \phi(y))$, T de paramètres $(a', b', \alpha', \beta', F)$, avec $x = ca', \phi(x) = \gamma\beta', \dots$ est attestée. La même indétermination prévaut pour les transformations initiales.*

Cette propriété signifie précisément que l'observation d'un couple de paires alternantes atteste simultanément plusieurs transformations : il y a donc indétermination sur la nature de la transformation extraite.

Reportons nous à nouveau à la figure 4.8 : nous avons mentionné que le système représenté permet d'attester la transformation initiale ($a \rightarrow fa, /a/ \rightarrow /fa/$). Nous aurions aussi bien pu insister sur la transformation initiale ($ac \rightarrow fac, /ak/ \rightarrow /fak/$), ou encore sur ($acte \rightarrow facte, /a/ \rightarrow /fa/$)...

On fera référence à la transformation initiale attestée par (x, y) telle que $a = pgcp(x, y)^{-1}x, \alpha = pgcp(\phi(x), \phi(y))^{-1}\phi(x), \dots$ comme la transformation initiale *minimale* attestée par (x, y) . Une définition similaire permet de parler de transformation minimale finale. Les quatre transformations minimales de notre exemple de référence sont définies par ($\lambda \rightarrow f, \lambda \rightarrow /f/$), et son inverse pour ce qui est des transformations initiales minimales; et ($ion \rightarrow eur, /sjõ/ \rightarrow /tœʁ/$) et son inverse pour ce qui est des transformations minimales finales.

Propriété 4.4.5 (Stabilité par composition) *L'ensemble des transformations initiales (resp. finales) est stable pour la loi de composition des transformations. Par contre, la réunion de $\mathcal{T}_I$ et de $\mathcal{T}_F$ n'est pas stable par composition.*

On notera par contre que la propriété de stabilité ne s'étend pas aux sous-ensembles de $\mathcal{T}_I$ et de $\mathcal{T}_F$ correspondant aux seules transformations attestées ou paradigmaticques sur $\mathcal{L}$.

4.4 Prononcer par application de transductions paradigmatiques 169

D'un point de vue formel, on notera que, si l'on se restreint aux transformations dans le domaine graphémique, alors l'ensemble des transformations graphiques paradigmatiques minimales sur $\mathcal{L}$ permet de définir un mécanisme de *calcul associatif*, sous la forme d'un système de relations de Thue (Gross & Lentin 1970).

Évaluer une transformation graphique Si l'on se restreint maintenant aux transformations graphiques paradigmatiques, on associe à chaque transformation la note suivante :

$$S(T_G) = \frac{|\{x \in D_{T_G} \text{ et } T_G(x) \in \mathcal{L}\}|}{|D_{T_G}|} \quad (4.1)$$

La note définie à l'équation 4.1 peut s'interpréter en termes probabilistes comme suit. Dans la mesure où l'on va chercher à construire des prononciations pour des chaînes déterminées, tous nos calculs se feront à x fixé. En particulier on ne s'intéressera qu'aux transformations qui sont définies pour x . $S(T_G)$ s'interprète alors comme une estimation de la probabilité, conditionnelle à l'appartenance de x à D_{T_G} , du succès de la transformation. Pour que cette estimation soit fiable, on suppose que les mots inconnus ressemblent suffisamment aux mots du lexique d'apprentissage. Nous verrons à la section 4.4.2 comment cette note a été utilisée.

Définitions finales On appellera transformation *libre* une transformation dont le domaine de définition est $\mathcal{G}$ tout entier; i.e. qui est de la forme: $x \rightarrow ax$ ou $x \rightarrow xa$. Ces transformations posent un problème algorithmique particulier, car il est possible de les appliquer *ad nauseam* à une forme quelconque. La possibilité d'extraire de telles transformations d'un lexique correspond à deux types de faits linguistiques bien distincts :

- d'une part l'existence de paires minimales graphémiques dont l'un des termes est nul : ainsi le couple *action-faction*, témoin de l'alternance λ -f;
- d'autre part, l'existence de paires de termes dont l'un dérive morphologiquement de l'autre par stricte affixation : par exemple la paire *donner-redonner*.

Il ne fait pas de doute que le premier type d'alternance est phonologiquement conditionné par des principes phonotactiques : il n'existe, par exemple, pas de paire minimale attestant l'alternance λ -f et dont le premier terme commence par une obstruante. En ce qui concerne le second type, l'existence d'un conditionnement phonologique est moins systématique.

En conséquence, pour caractériser ce type de transformations plus complètement, nous avons été amené à introduire la notion de *contexte apparent* (C_a) d'une transformation graphémique libre⁹ T_G de la manière suivante :

9. La tentation était grande de généraliser cette notion à toutes les transformations. Un des avantages immédiats aurait été de mieux spécifier les domaines de validité des transformations

$$C_a(T_G^I) = \{c \in G \text{ tels que } \exists x \in \mathcal{L}, T_G(x) \in \mathcal{L} \text{ et } x = cy\}$$

$$C_a(T_G^F) = \{c \in G \text{ tels que } \exists x \in \mathcal{L}, T_G(x) \in \mathcal{L} \text{ et } x = yc\}$$

Notons que cette caractérisation par un contexte graphémique de taille fixe et limité ne permet de saisir que très imparfaitement les conditionnements réels qui contraignent les transformations libres, en particulier pour celles qui sont la trace de phénomènes morphologiques dérivationnels. Cela est tout spécialement flagrant des conditionnements morpho-syntaxiques (c.f. la section 4.4.1) : enregistrer la liste des caractères devant lesquels la préfixation de *dé* (comme dans *faire-défaire*) a été observée ne présente que peu d'intérêt, puisque ce qui conditionne la transformation est la catégorie morpho-syntaxique de la lexie à transformer ! Cette vision très simpliste du contexte gagnerait largement à être améliorée, en particulier pour éviter d'engendrer pendant la recherche un trop grand nombre de formes improbables. Nous ne l'avons pas fait, pour rester fidèle au principe qui permet de comparer les différentes approches du problème de la transcription, et qui exige que les seules données utilisées soient les chaînes graphiques et phonétiques.

La représentation strictement extensive (i.e. sous la forme d'une liste de caractères) de ces contextes limite également leur validité, puisqu'elle ne permet pas de se prémunir contre la présence de couples exceptionnels. Considérons en effet l'exemple suivant : l'adjonction du préfixe /in/ est, en anglais (comme en français), phonologiquement conditionnée par la nature de la consonne initiale du mot préfixé. Si c'est une bilabiale (/mbp/), l'allomorphe préfixé sera /im/, sinon, au cas des liquides (/lʃr/) près, ce sera /in/. Ce phénomène se traduit dans les formes graphémiques par la préfixation de *im* et de *in* respectivement. Nous devrions donc enregistrer deux transformations, l'une correspondant à la préfixation de *im*, de contexte égal à {*mpb*}, l'autre correspondant à la préfixation de *in*, avec un contexte égal aux autres graphèmes. L'existence accidentelle d'une paire telle que *put-input* dans le lexique NETTALK, contraint l'inclusion du *p* dans le contexte de la préfixation de *in*, et enlève toute pertinence à la définition du contexte apparent de la préfixation de *in*.

Extraction de couplages de transformations

Recensement des transformations graphiques La première étape de notre mécanisme consiste à extraire du lexique d'apprentissage toutes les transformations graphiques paradigmatiques. Pour ce faire, nous avons utilisé la même procédure que celle mise en œuvre dans PPP, en l'appliquant à la liste des

extraites, qui ne sont qu'imparfaitement contraintes par le principe de minimalité des transformations. Ainsi, la nominalisation des adjectifs en *-able*, sur le schéma *solvable-solvabilité* se manifeste au niveau graphique par la transformation *le* → *ilité*. Cette transformation n'est effectivement productive que pour des formes qui se terminent par *able* (ou *ible*). La démarche ici présentée s'interdit d'enregistrer cette information (importante) supplémentaire ; en conséquence la transformation précédente sera appliquée à toute forme se terminant par *le* (par exemple *fil*, ou *cannelle*). La motivation de ce choix réside, pour une part, dans la volonté de conserver à notre mécanisme des capacités suffisantes de généralisation, auxquelles le respect systématique de contraintes contextuelles serait nuisible ; et, pour une autre part, dans la remarque que ce type de régularité peut en fait être saisi d'une autre manière, au niveau phonétique (c.f. le dernier paragraphe de la section suivante).

4.4 Prononcer par application de transductions paradigmatiques 171

lexies du lexique d'exemples. Ainsi, pour chaque mot connu, nous avons énuméré toutes les transformations graphiques paradigmatiques qui pouvaient lui être appliquées de façon à en dériver un autre élément du lexique. Cette démarche s'exprime formellement par l'algorithme suivant (algorithme 5), qui est appliqué à chaque mot du lexique d'apprentissage.

Algorithme 5: Extraction de transformations graphémiques paradigmatiques

Entrée: Une chaîne x

Sortie: Un ensemble de transformations $\mathcal{T} = \{T_i\}$, avec :

$\forall i, T_i(x) \in \mathcal{L}$

PCP-EXTRACTION(x)

- (1) $D(x) \leftarrow \{y \in \mathcal{L} / x > y \text{ et } (t = pgcp(x, y)) \neq \lambda\}$
- (2) **pour tout** $y \in D(x)$
- (3) $P(x, y) \leftarrow \{(z, t) \in \mathcal{L} \times \mathcal{L} / y > t \text{ et } z = T_{xy}^F(t)\}$
- (4) **si** $P(x, y) \neq \emptyset$
- (5) Enregistrer les transformations T_{xy}^F et T_{xt}^I
- (6) **fin**

Dans la pratique, les propriétés de symétrie et de dualité permettent d'accélérer cette étape d'extraction : c'est le sens des conditions d'ordonnancement (à comprendre ici au sens de l'ordre lexicographique) des lignes 1 et 3. De plus, dans la mesure où cette étape ne comporte aucun dénombrement, il est possible d'en réduire la durée en ne cherchant les transformations qu'à partir d'un petit sous-ensemble des lexies disponibles. Dans les expériences que nous avons menées, nous avons appliqué l'algorithme d'extraction à des échantillons aléatoires représentant entre 10 et 20 % des exemples disponibles.

Couplage et évaluation Une fois ce recensement terminé, nous avons évalué chaque transformation graphique, en calculant la note associée, ainsi que, dans le cas des transformations libres, le contexte apparent. Cette étape d'évaluation permet également d'identifier, transformation graphique par transformation graphique, les différentes transformations phonétiques associées. Durant cette étape, on calcule également une note pour les transformations phonétiques selon :

$$S(T) = \frac{|\{x \in D_{T_G} \text{ et } T_G(x) \in \mathcal{L} \text{ et } \phi(x) \in D_{T_P} \text{ et } T_P(\phi(x)) = \phi(T_G(x))\}|}{|\{x \in D_{T_G} \text{ et } T_G(x) \in \mathcal{L} \text{ et } \phi(x) \in D_{T_P}\}|} \quad (4.2)$$

Cette note s'interprète également comme un estimateur de la probabilité, conditionnelle à l'appartenance de x et de $\phi(x)$ aux domaines de définition respectifs de T_G et de T_P , que T_P commute avec ϕ .

Si on ne se fonde que sur cette note, la comparaison des notes associés à deux transformations phonétiques pose toutefois un problème délicat. Avant d'aborder ce problème, et de présenter la manière dont nous l'avons résolu, il apparaît nécessaire de se pencher sur un exemple précis de transformation que nous avons pu extraire durant nos expériences.

T_G	x	$\rightarrow$	$x\text{-ly}$		0,005	0,413
T_{P_1}	x	$\rightarrow$	$x\text{-/li:/}$	<i>good</i>	0,681	0,756
T_{P_2}	$x\text{-/t/}$	$\rightarrow$	$x\text{-/ədli:/}$	<i>marked</i>	0,1	0,071
T_{P_3}	$x\text{-/əl/}$	$\rightarrow$	$x\text{-/əli:/}$	<i>equal</i>	0,571	0,8
T_{P_4}	$x\text{-/əl/}$	$\rightarrow$	$x\text{-/li:/}$	<i>capable</i>	0,619	0,158
T_{P_5}	x	$\rightarrow$	$x\text{-/i:/}$	<i>cool</i>	0,769	0,777
T_{P_6}	$x\text{-/i:n/}$	$\rightarrow$	$x\text{-/enli:/}$	<i>ear</i>	0,5	1
T_{P_7}	$x\text{-/ɪd/}$	$\rightarrow$	$x\text{-/aɪdli:/}$	<i>id</i>	1	1
T_{P_8}	$x\text{-/ɪv/}$	$\rightarrow$	$x\text{-/aɪdli:/}$	<i>live</i>	1	0,5
T_{P_9}	$x\text{-/θ/}$	$\rightarrow$	$x\text{-/ðli:/}$	<i>loath</i>	1	1
$T_{P_{10}}$	x	$\rightarrow$	$x\text{-/laɪ/}$	<i>imp</i>	0,011	1
$T_{P_{11}}$	$x\text{-/ɪr/}$	$\rightarrow$	$x\text{-/ɜli:/}$	<i>ear</i>	1	1
$T_{P_{12}}$	$x\text{-/ɔn/}$	$\rightarrow$	$x\text{-/onli:/}$	<i>on</i>	1	0,142

TAB. 4.3 – Un exemple de transformation finale.

Quelques co-transformations Le résultat final de cette étape d'apprentissage est un ensemble de transformations graphiques informées notamment par les couplages avec une série de transformations phonétiques : on dispose en fait d'un ensemble de co-transformations. Un exemple réel de transformations graphiques extraites de nos expériences avec le lexique NETTALK, ainsi que les transformations phonétiques et scores afférents, est donné dans la table 4.3. Cet exemple correspond approximativement à la formation en anglais d'adverbes à partir d'adjectifs par suffixation de *-ly*. Pour chaque transformation, nous donnons dans la table un exemple, correspondant à la forme graphique originale, ainsi que les scores de la transformation et de la transformation inverse.

Quelques remarques concernant ce tableau. Notons tout d'abord que la transformation phonétique T_{P_3} est, malgré les apparences, minimale : le système de transcription phonétique utilisé dans NETTALK distingue en effet le /l/ vocalique (qui termine la transcription de *equal*) de la séquence /xl/ (que l'on retrouve dans *equally*).

On constate, par ailleurs, que certaines transformations phonétiques, ou leur inverse, ont un score égal à 1. Ces transformations sont numérotées de T_{P_6} à $T_{P_{12}}$. Ce sont en fait des transformations qui ne sont attestées que pour un seul couple de mots. Afin de rester en cohérence avec le principe d'extraction de transformations *paradigmatiques*, nous avons dans les faits purement et simplement ignoré ces transformations-là. On notera que pour cet exemple, le principe paradigmatique permet de se débarrasser des transformations phonétiques les plus pathologiques : ainsi $T_{P_{10}}$, improprement extraite à partir de la paire *imp-implicitly*, dont on ne peut dire quelle s'inscrit dans ce paradigme dérivationnel d'adverbialisation (*imp* est un nom, et *imply* un verbe !). On pourrait en dire autant de $T_{P_{11}}$, uniquement attestée par la paire *ear-early*, dont le premier terme est un nom ; ou de T_{P_8} , attestée par la paire *live* (dans son emploi verbal)-*lively*.

Les transformations T_{P_1} et T_{P_5} partagent la caractéristique d'avoir un domaine de définition non-contraint, et sont donc applicables à n'importe quelle chaîne phonétique. Elles se distinguent sous ce rapport de T_{P_2} , T_{P_3} et T_{P_4} , qui spécifient des domaines de définition plus restreints. Ceci met en lumière un fait important de notre démarche : les transformations graphémiques sont

4.4 Prononcer par application de transductions paradigmatiques 173

T_{G_1}	x	$\rightarrow$	$t-x$	17940	111
$T_{P_{11}}$	x	$\rightarrow$	$/t/-x$	111	68
$T_{P_{12}}$	$/h/-x$	$\rightarrow$	$/ð/-x$	21	4
$T_{P_{13}}$	$/h/-x$	$\rightarrow$	$/θ/-x$	21	12
$\vdots$			$\vdots$	$\vdots$	$\vdots$
T_{G_2}	$x-ion$	$\rightarrow$	$x-e$	885	250
$T_{P_{21}}$	$x-/fən/$	$\rightarrow$	$x-/t/$	230	189
$T_{P_{22}}$	$x-/eʃən/$	$\rightarrow$	$x-/ət/$	195	9
$\vdots$			$\vdots$	$\vdots$	$\vdots$

TAB. 4.4 – Quelques transformations problématiques

(aux transformations libres près) extraites et utilisées comme des règles de production indépendantes du contexte (graphémique). La donnée des domaines de définition des transformations phonétiques associées va toutefois aider à imposer, d'une certaine manière, un contexte (phonémique) *a posteriori* : pour qu'une transformation soit utilisable pendant la construction d'une solution, il faudra en effet (c.f. la section 4.4.2) que les domaines de définition des transformations phonétiques soient compatibles avec les chaînes à récrire.

Pour compléter notre description de cet ensemble de transformations, notons que le contexte graphémique apparent associé est : $\{bdeghklmnp\}$, reflétant la relative indépendance vis-à-vis du contexte graphique de cette transformation; nous l'avons vu, le conditionnement réel de cette transformation devrait être représenté par des contraintes sur les catégories morpho-syntaxiques de départ et d'arrivée.

Conflits entre transformations phonétiques L'étude de l'exemple de la transformation graphique $x \rightarrow x\text{-}ly$ permet finalement de mettre en évidence le fait que certaines transformations phonétiques (T_{P_i} par exemple) ont un domaine de définition strictement plus large que d'autres. Ce point, en relation avec l'interprétation en termes de contexte des domaines de définition, nous confronte au problème suivant, exemplifié au travers de deux autres transformations (voir table 4.4) issues du traitement du lexique NETTALK. On a fait figurer sur cette table en lieu et place des scores le cardinal du domaine de définition conditionnel, ainsi que le nombre d'éléments de ce domaine pour lesquels la transformation est attestée.

Cette table présente une sous-partie des transformations phonétiques associées respectivement aux transformations graphiques $\lambda \rightarrow t$ et $ion \rightarrow e$: la première reflète l'existence de nombreuses paires minimales graphiques $\lambda-t$, telles que *ale-tale*, *error-terror*, ou encore *race-trace*; la seconde correspond approximativement au schéma général inverse de la nominalisation en anglais des verbes de racine latine, schéma attesté par les paires *activation-activate*, *creation-create*, ou encore *federation-federate*.

Dans le premier cas, le rôle contextuel joué par le domaine de définition est clairement mis en évidence : le corrélat phonétique le plus fréquemment attesté de la transformation est $\lambda \rightarrow /t/$, soit une transformation dont le domaine de

définition est non-contraint. Il est clair, toutefois, que pour tous les mots qui commencent par un *h* graphique, cette règle est systématiquement prise en défaut, puisque le terme résultant commencera par le digraphe *th*, qui se lit en anglais exclusivement /θ/ ou /ð/. C'est le sens à donner aux transformations $T_{P_{12}}$ et $T_{P_{13}}$, dont les domaines de définition spécifient précisément les mots qui commencent par un *h* graphique (et donc un /h/ phonétique), et qui sont attestées par des paires telles que *his-this*, *hat-that*, ou encore *high-thigh*. Nous sommes ici dans une situation où la transformation la plus générale ($T_{P_{11}}$) n'est en fait attestée que sur un sous-ensemble de son domaine de définition disjoint du domaine de définition de $T_{P_{12}}$ et de $T_{P_{13}}$, à savoir l'ensemble des chaînes de phonèmes qui ne commencent pas par un /h/. Pour utiliser cette information lors de l'étape de construction des transcriptions, il ne faudrait alors l'appliquer que dans les cas où les transformations plus spécifiques sont inopérantes, et ce indépendamment des indices de confiance numériques associés aux transformations qui, dans ce cas précis, conduiraient à privilégier $T_{P_{11}}$.

La transformation $\text{ion} \rightarrow \text{e}$ nous propose un exemple exactement inverse du précédent. Ici encore, une transformation phonétique ($T_{P_{21}}$) possède un domaine de définition plus spécifique qu'une transformation concurrente. Il se trouve toutefois que dans ce cas, la transformation la plus générale est celle qui est majoritairement attestée y compris sur le domaine de définition de $T_{P_{22}}$, c'est-à-dire l'ensemble des chaînes phonémiques se terminant en /eʃən/. Pour toutes ces chaînes en effet, la transformation majoritairement attestée est /eʃən/ $\rightarrow$ /et/, interprétée minimalement en $T_{P_{21}}$. Cette transformation n'est invalidée que par une dizaine de cas, qui d'ailleurs ne correspondent pas à des transformations Nom-Verbe, mais à des transformations Nom-Nom¹⁰. Ces quelques exemples ont suffi à attester la transformation $T_{P_{22}}$. Suivre la stratégie que l'exemple précédent nous suggère, à savoir n'appliquer les transformations les plus générales que s'il n'en existe pas de plus spécifique, est ici suicidaire, puisque cela conduit à appliquer dans tous les cas une transformation hautement irrégulière, au détriment de la règle générale.

Pour résoudre ce dilemme, nous avons adopté l'heuristique suivante, qui se glose en « ignorer toutes les co-transformations dont la partie phonétique n'est pas majoritairement attestée sur leur domaine de définition ». L'ensemble des transformations que nous considérerons dans la suite de notre description est alors un sous-ensemble de $\mathcal{T}_L$ qui vérifie la propriété suivante :

$$\hat{\mathcal{T}}_L = \{T \in \mathcal{T}_L, \forall T' / D_{T'} \supset D_T, |L_T| \geq |L_{T'} \cap D_T|\} \quad (4.3)$$

Ceci nous permet de filtrer correctement des transformations telles que $T_{P_{22}}$, tout en conservant $T_{P_{12}}$ et $T_{P_{13}}$. Nous pourrions ainsi conserver au domaine de définition son interprétation contextuelle lors de l'étape de transcription pro-

10. Il est connu qu'en anglais, une forme de nominalisation des verbes relativement productive laisse inchangées les formes graphiques, le déplacement de l'accent tonique étant la seule marque de la transformation. Par exemple, la paire *record:V-record:N* ne se distingue que par la position de l'accent, positionné sur la dernière syllabe dans le cas du verbe, sur la première dans le cas du nom. Le lexique NETTALK contient quelques noms en -ate (*graduate*, *delegate*) dérivés de verbes suivant ce schéma, qui se distinguent phonétiquement des verbes par leur finale en *ət* (à la place de *et*), puisque la désaccentuation de la voyelle finale causée par la nominalisation a entraîné également une modification de sa qualité.

prement dite (voir section 4.4.2).

Cheminement à l’aveugle

Lorsqu’un mot de prononciation inconnue est proposé au système, l’idée générale, rappelons-le, consiste à lui appliquer des transformations graphiques jusqu’à reconstruction d’un *analogue*, puis appliquer à la prononciation de cet analogue la série de transformations phonétiques inverse. Nous explicitons dans cette section les principes de sélection des transformations à appliquer au mot cible pour en dériver un terme du lexique. Nous illustrerons les principes de recherche mis en œuvre par l’exemple suivant. Le mot cible à prononcer est le mot *lecteur*, et les transformations graphiques disponibles sont celles qui sont listées dans la table 4.5.

Les premières ramifications de l’arbre des dérivations du mot cible compte-tenu de l’ensemble des transformations sont représentées sur la figure 4.10, où l’on a arbitrairement numéroté les nœuds. On a fait l’hypothèse additionnelle que la transformation graphique correspondant à la préfixation de *f* est suffisamment contrainte par son contexte apparent pour que des dérivés tels que *fsection* ne soient pas engendrés.

Pour compléter la description de cet exemple, disons finalement que l’analogue à atteindre est *section* (entouré sur le graphique).

Tâtonnements L’étape de recherche se déroule schématiquement de la manière suivante : le mot cible est tout d’abord cherché dans le lexique. S’il n’y est pas trouvé, alors on lui applique toutes les transformations possibles, pour obtenir une liste de dérivés. Le même procédé est alors appliqué récursivement à chacun des dérivés ainsi construit, jusqu’à ce que l’on aboutisse à un mot du lexique.

Ainsi exprimée toutefois, cette approche cumule un certain nombre d’inconvénients pratiques. Il lui manque tout d’abord une condition d’arrêt, puisque pour une cible, un lexique, et un jeu de transformations donnés, rien ne garanti l’existence d’un dérivé dans lexique. Un autre handicap majeur de cette stratégie est que la liste des dérivés croît de manière explosive dès que le jeu de transformations est un tant soit peu étoffé, puisque chaque mot évalué conduit à la création de $|\{T, x \in D_T\}|$ termes dérivés (de l’ordre de quelques centaines, dans les tests conduits sur le lexique NETTALK).

Tâtonnements et analyse syntaxique Il est possible de reformuler le problème de la recherche d’un analogue comme un problème d’analyse syntaxique,

Transformations initiales			Transformations finales		
<i>f</i>	$\longleftrightarrow$	λ	<i>eur</i>	$\longleftrightarrow$	<i>ion</i>
<i>f</i>	$\longleftrightarrow$	<i>r</i>	<i>e</i>	$\longleftrightarrow$	
<i>s</i>	$\longleftrightarrow$	<i>l</i>	<i>r</i>	$\longleftrightarrow$	<i>se</i>

TAB. 4.5 – Un ensemble de transformations pour lire *lecteur*.

4.4 Prononcer par application de transductions paradigmatiques 177

en considérant la grammaire dont les caractéristiques sont les suivantes :

- les symboles terminaux sont les symboles alphabétiques;
- le seul symbole non-terminal est S
- les règles de production se composent pour une part de règles décrivant intégralement le lexique sous la forme $S \rightarrow x$, avec $x \in \mathcal{L}$; et contiennent d'autre part toutes les transformations extraites durant l'apprentissage.

Trouver un analogue à un mot cible revient dans ce contexte exactement à produire une analyse du mot dans la grammaire, c'est-à-dire à trouver la série de règles permettant de produire S , production qui n'est possible que si l'on a atteint un mot du lexique. Compte-tenu malheureusement de la forme absolument quelconque des récritures qui sont extraites, pour lesquelles les parties droite et gauche sont de taille variable (en particulier elles peuvent être vides) la grammaire ainsi spécifiée est du type le plus général (grammaire à structure de phrase de type 0). Pour cette classe de grammaires il n'existe pas d'algorithme général permettant d'effectuer une analyse syntaxique; par ailleurs les techniques d'analyse disponibles pour certaines sous-classes particulières sont toutes exponentielles (Grune & Jacob 1990).

Ces résultats négatifs nous ont conduit d'une part à introduire un certain nombre d'heuristiques et de filtres pour limiter autant que possible la croissance de l'arbre à explorer; d'autre part à élaborer deux stratégies d'exploration en relation avec des conditions d'arrêt spécifiques. Ces deux aspects importants de notre travail sont abordés tour à tour dans les sections qui suivent. Il est certain qu'un travail important reste à accomplir pour améliorer le mécanisme de recherche. Les principes parfois *ad-hoc* que nous avons adoptés nous ont néanmoins permis de confronter notre intuition à des lexiques réels, et de produire un certain nombre de résultats préliminaires validant la pertinence de notre approche.

Heuristiques d'élagage

Si la fixation de conditions d'arrêt explicites garantit que nos algorithmes s'arrêtent effectivement, en l'état ils sont excessivement gourmands en temps de calcul. Un des problèmes typiques auquel cette approche nous confronte est l'existence de dérivation qui s'enchaînent de manière cyclique, conduisant à la construction répétée (et à l'évaluation qui s'en suit) des mêmes candidats-analogues. Ceci s'observe sur le graphique 4.10, sur lequel on observe la répétition des mêmes candidats en divers endroits de l'arbre. Ainsi, *flecteuse* apparaît aux nœuds (2.3) et (4.2) de la dérivation.

Traitement des cycles L'exemple caricatural de tels cycles est l'application successive d'une transformation et de son inverse, qui sont, par principe, toujours simultanément attestées. Si de tels cycles sont relativement faciles à détecter, d'autres, par exemple ceux qui mettent en jeu des échanges répétés de consonnes initiales, et qui peuvent être de taille quelconque, le sont beaucoup moins.

De manière à limiter les effets néfastes de l'enchaînement cyclique de ces dérivations, nous avons été conduit à introduire les deux principes suivants :

Heuristique 4.4.1 *une transformation initiale (resp. finale) ne peut s'appliquer si la chaîne qu'elle récrit résulte de l'application de la transformation initiale (resp. finale) qui précède dans la dérivation.*

Heuristique 4.4.2 *Il est interdit d'appliquer deux fois la même transformation au même dérivé au cours d'une dérivation.*

Le premier principe est relativement ad-hoc, et repose sur l'observation que, malgré la non-stabilité par composition de l'ensemble des transformations attestées sur $\mathcal{L}$, si les transformations $ax \rightarrow bx$ et $bx \rightarrow cx$ sont attestées, alors $ax \rightarrow cx$ l'est également, bien qu'avec une note différente du produit des notes des deux autres. En interdisant les enchaînements de ce type, le composé cx ne pourra dériver que de ax que par application d'une transformation, et non pas d'une série de transformations. Notons que ce principe interdit de fait la création de cycles de taille 2 (une transformation suivie de son inverse).

L'action de ce principe se manifeste sur le graphe 4.10 par la suppression d'un certain nombre de ramifications: en plus des multiples occurrences du mot cible (nœuds (1.1), (2.1), ...), un candidat tel que *rlecteur* (nœud (2.4)) ne sera plus engendré, puisqu'il résulte d'un échange du *f* de *flecteur* par un *r*, qui lui-même provient d'une réécriture du *f*.

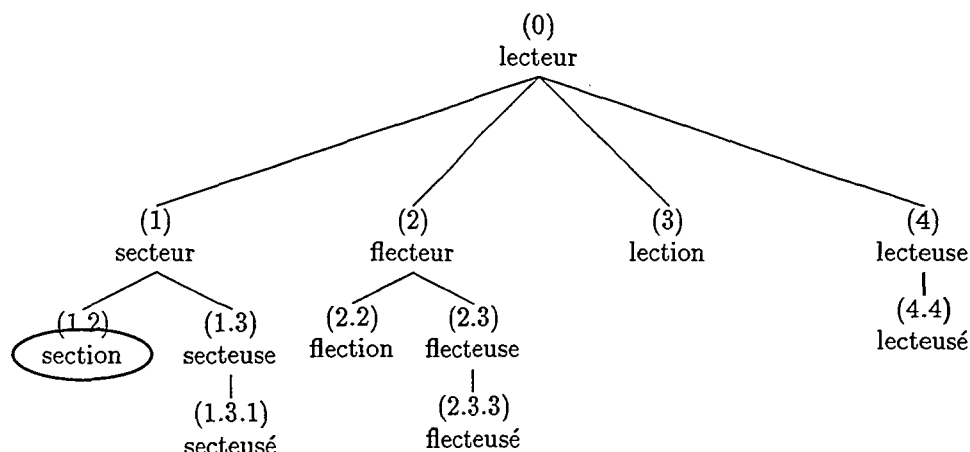
Le second principe interdit de même qu'un même candidat formé par l'application de deux suites de dérivations différentes engendre plusieurs fois la même descendance. Il garantit ainsi que si un cycle de transformation conduit un même candidat à apparaître deux fois au cours d'une même dérivation, alors la ré-application de la première transformation du cycle est bloquée.

Les effets de cette heuristique ne sont pas visibles sur notre exemple de la page 175. Pour qu'ils le soient, il faudrait, par exemple, que la transformation *euse* → *ion* soit attestée. La seconde heuristique bloque dans ce cas des dérivations cycliques plus complexes telles que : *lection* → *lecteur* → *lecteuse* → *lection* → *lecteur*, qui ne sont pas prises en compte par l'heuristique 4.4.1.

Commutation Un second problème majeur provient du fait que transformations initiales et finales commutent entre elles sur des sous-parties de leur domaine de définition. Soient en effet T_G^I définie par $ax \rightarrow bx$ et T_G^F définie par $xu \rightarrow xv$, ces deux transformations commutent pour tous les éléments de l'intersection de leur domaine de définition qui sont de la forme azu . Pour ces chaînes, le dérivé bzv sera construit par applications successives dans un ordre quelconque de T_G^F et de T_G^I , et sera donc, ainsi que sa descendance, évalué deux fois.

Pour limiter ce phénomène, nous avons adopté le principe suivant :

Heuristique 4.4.3 *Si une transformation initiale s'applique à un terme dérivé construit par application d'une transformation finale, et si ces deux transformations commutent, alors la dérivation est bloquée.*

FIG. 4.11 – L'arbre des dérivés du mot cible *lecteur*, après élagage.

Par ce biais, les successions transformation initiale-transformation finale ne peuvent s'appliquer au cours d'une dérivation que dans cet ordre-là. Cette heuristique permet d'éviter la double génération d'un candidat tel que *secteuse*, qui apparaît deux fois sur le graphique 4.10. La transformation initiale $s \rightarrow f$ qui engendre le nœud (4.3) est en effet bloquée, puisqu'elle commute avec la transformation qui échange les suffixes *eur* et *ion*.

Transformations libres Nous avons pris soin, durant la période d'extraction des transformations, d'enregistrer les contextes d'application des transformations libres. Ces contextes sont utilisés durant les dérivations pour restreindre les domaines d'application.

Filtrage des transformations Un troisième moyen plus radical pour contrôler la croissance de l'arbre de dérivation consiste tout simplement à limiter l'ensemble des transformations utilisées pour la construction de dérivés. Les filtres que nous avons mis en place concernent à la fois la taille maximale des chaînes réécrites par une transformation, et la fréquence observée des transformations.

Le résultat global de l'application cet ensemble d'heuristiques est représenté sur la figure 4.11, qui reproduit *intégralement* l'arbre des dérivés de *lecteur*, compte-tenu des transformations listées dans la table 4.5.

Exploration de l'arbre

Gestion de la pile Le mécanisme général d'exploration de l'arbre est implanté par le biais de la gestion d'une pile de candidats à l'analogie, i.e. de chaînes qui sont potentiellement des lexies. Cette pile est initialisée en empilant le mot cible. Chaque pas dans l'exploration de l'arbre correspond à quatre

opérations distinctes :

- dépilage du candidat-analogue se trouve en tête de la liste;
- recherche dans le lexique de ce candidat-analogue. S'il est trouvé, l'exploration s'arrête, et on passe à l'étape de construction de la prononciation par cheminement inverse;
- génération de tous les dérivés possibles du candidat-analogue en cours d'examen. Chaque dérivé est nanti d'une note, et est ajouté en queue de la pile;
- reclassement de la pile par ordre de notes décroissantes.

Ces quatre étapes constituent *l'évaluation* d'un candidat-analogue. La stratégie d'exploration de l'arbre qui résulte de cette procédure correspond donc à un parcours en profondeur d'abord.

L'ordre selon lequel l'arbre des dérivations est parcouru est bien entendu crucial pour les performances de l'étape de recherche, qu'on les mesure en fonction du temps nécessaire pour atteindre un mot du lexique ou en termes de la ressemblance entre le mot inconnu et l'analogue éventuellement exhibé. Le choix de la fonction de notation est donc déterminante pour le succès de cette stratégie. Nous avons identifié et implanté deux critères de classement différents, qui correspondent chacun à une manière différente de considérer les analogues potentiels.

Du double visage des candidats Il serait en effet d'un côté souhaitable de pouvoir ordonner les dérivés selon leur chance *a priori* d'être effectivement dans le lexique, de façon à limiter le nombre de recherches nécessaires. Mais si la recherche échoue, il est par contre souhaitable de commencer par engendrer en priorité les dérivés des termes pour lesquels la probabilité qu'un dérivé soit dans le lexique est forte. Autrement dit, il existe deux principes de classement des dérivés : un qui les ordonne selon la probabilité qu'ils soient dans le lexique, un qui les ordonne selon la probabilité qu'ils aient des descendants directs parmi les lexies.

Malheureusement, ces deux principes sont parfois contradictoires, comme en témoigne l'exemple (artificiel) suivant. De nombreuses transformations graphiques très productives en français consistent à échanger deux consonnes initiales. Un exemple est $p \leftrightarrow r$, attesté par les paires *page-rage*, *paie-raie*, ou encore *poule-roule*. Supposons que le mot cible soit *plage*. Cette chaîne est dans le domaine de définition de la transformation; de plus, son application a *a priori* (les transformations sont non-contextuelles) de bonnes chances de produire un analogue. Il est donc souhaitable que, parmi tous les dérivés de *plage*, *rlage* soit évalué assez tôt. Ce dérivé présente toutefois une séquence initiale *rl* invalide : ses descendants directs par transformation finale ont en conséquence peu de chance d'être trouvés dans le lexique. Ce dernier point milite pour que l'on engendre que tardivement les dérivés de *rlage*, afin d'éviter d'encombrer la pile par des candidats improbables.

4.4 Prononcer par application de transductions paradigmatiques 181

Les deux principes d'exploration que nous avons mis en œuvre, et qui sont décrits dans les paragraphes suivants, correspondent grossièrement à favoriser l'un ou l'autre de ces aspects dans l'ordre d'évaluation.

Chercher à bon escient Le premier type de cheminement proposé repose sur l'interprétation probabiliste des scores associés aux transformations. Pour un mot x du lexique dans le domaine de définition de T_G , $S(T_G)$ estime la probabilité pour que $T_G(x)$ soit également un mot du lexique. Si l'on fait l'hypothèse que les mots à transcrire sont identiques aux mots appris (et ont donc une probabilité 1 d'être dans le lexique), alors on peut calculer, à chaque étape de la dérivation, la probabilité qu'un dérivé $y = T_n o \dots o T_1(x)$ de x soit dans le lexique par : $\prod_{i=1}^n S(T_i)$.

En utilisant cette valeur comme la note du dérivé dans la pile, on obtient un principe d'ordonnancement des candidats qui conduit évaluer en priorité les dérivés qui sont le plus probablement dans le lexique; et à minimiser en principe le nombre de recherches qui vont être effectuées.

Si l'on s'assure de plus qu'aucune transformation n'a une probabilité de succès rigoureusement égale à 1, alors la probabilité qu'un mot soit dans le lexique décroît strictement au cours des dérivations. Ceci permet d'utiliser de manière naturelle le critère d'arrêt suivant : stopper la recherche dès que tous les candidats ont une probabilité inférieure à un seuil donné d'être dans le lexique.

Il est important de remarquer que, dans la même optique, il aurait été possible de faire également intervenir d'autres mesures de la probabilité qu'un mot appartienne au lexique; par exemple des modèles de mots à base de n -grammes permettant de rendre compte de la satisfaction par la chaîne dérivée de contraintes phonotactiques.

Engendrer à bon escient L'approche alternative se fonde sur la remarque suivante : si x est un candidat-analogue, $E(x) = \sum_{x \in D_T} S(T)^{11}$ correspond à une estimation de l'espérance du nombre de dérivés de x dans le lexique. En ordonnant les candidats en fonction de $E(x)$, on oriente l'exploration de l'arbre dans le sens de l'évaluation prioritaire de ceux qui sont le plus susceptibles d'avoir un dérivé dans le lexique.

Adopter ce point de vue nécessite toutefois de changer l'ordre dans lequel les opérations d'évaluation sont conduites, et, en particulier, impose de chercher les candidats dans le lexique au moment où ils sont engendrés, et non pas au moment où ils sont évalués. En effet, la préférence attribuée à la dérivation des termes « prometteurs » suppose que l'on vérifie le succès de la dérivation le plus tôt possible. Sans cela, il est possible qu'un mot du lexique, en apparence peu prometteur, soit créé, mais plus jamais évalué. Une conséquence directe de ce réordonnancement des opérations est que plusieurs analogues sont susceptibles d'être exhibés de façon concurrente. Dans ce cas, tous les analogues contribuent

11. Le calcul est en fait légèrement plus complexe, et dépend de l'histoire des dérivations qui ont conduit à x , puisqu'au cours de la recherche les heuristiques d'élagages contribuent à réduire les transformations applicables à un candidat. La somme $\sum_{T, x \in D_T} S(T)$ se fait non plus sur l'ensemble de toutes les transformations paradigmatiques, mais sur l'ensemble de celles qui sont applicables, compte-tenu des heuristiques de filtrage.

Transformations initiales			Transformations finales		
λ	$\rightarrow$	f	0,001	eur	$\rightarrow$ ion 0,5
l	$\rightarrow$	s	0,05	e	$\rightarrow$ 0,05
				r	$\rightarrow$ se 0,1

TAB. 4.6 – Les probabilités des transformations permettant de lire lecteur.

à part égale à la construction de la prononciation, selon les modalités exprimées à la section 4.4.2. Par ce biais, nous sommes donc en mesure de reproduire des effets de « gang », caractéristiques des phénomènes analogiques observés dans les tâches de lecture.

Ce procédé fournit effectivement un début de solution au problème posé par des mots tels que *rlage* : un tel dérivé figurera mécaniquement dans le domaine de définition d'un nombre petit de transformations (en particulier de transformations initiales); il n'aura conséquemment en espérance que peu de descendants dans le lexique, et ne sera ainsi évalué que tardivement.

Il est toutefois important de noter que la note définie par $E(x)$ ne décroît pas pendant la dérivation, et ne peut en conséquence fonder un critère d'arrêt de la recherche. Nous avons dans le cadre de cette stratégie choisi de limiter *a priori* le nombre de dérivations possibles d'un mot (la profondeur de l'arbre des dérivations). La recherche s'arrête donc au bout d'un nombre fixé de dérivations.

Remarquons finalement que, dans le cas précis où la profondeur de la recherche est fixée à 1, et où toutes les transformations attestées sont conservées, cette stratégie est approximativement équivalente à celle mise en œuvre dans PPP.

Les stratégies de recherche Les deux stratégies de recherche et d'évaluations sont résumées formellement dans l'algorithme 6, sous les dénominations respectives PCP-RECHERCHE-1 pour la stratégie de recherche à bon escient, et PCP-RECHERCHE-2 pour la stratégie de génération à bon escient.

Nous ferons référence à ces stratégies sous les noms respectifs de SR_1 et de SR_2 dans la suite du texte.

L'exemple suivant, qui utilise à nouveau les transformations listées à la table 4.5 (page 175), illustre les différences entre ces deux stratégies. Nous avons assigné à des fins strictement illustratives les notes suivantes aux transformations utilisées (voir la table 4.6).

Sous ces hypothèses, on peut reconstruire précisément les deux parcours de l'arbre correspondant aux différentes stratégies de recherche. Ces parcours sont représentés, ainsi que la pile des candidats aux différentes étapes de l'algorithme dans la table 4.7, où chaque colonne du tableau correspond à une étape de l'exploration.

Dénombrons sur cet exemple le nombre d'opérations nécessaires pour trouver l'analogie en fonction de la stratégie choisie. Dans le cas SR_1 , 4 candidats sont évalués, correspondant à autant de recherches dans le lexique, et à 3 « générations » (dont une peu productive, puisque *lection* est sans descendant). Dans

Algorithme 6: Évaluation d'un candidat

Entrée: $\mathcal{T}$ un ensemble de transformations, P une pile de candidats, p un seuil d'arrêt

Sortie: un analogue, ou rien

PCP-RECHERCHE-1($\mathcal{T}, P, p$)

```
(1)  si  $x \in \mathcal{L}$  alors retourner  $x$ 
(2)   $P \leftarrow x$  avec  $n(x) = 0$ 
(3)  tant que  $P \neq \emptyset$  et  $s(t) \geq p$ 
(4)    dépiler  $t$ , de note  $s(t)$ 
(5)    si  $t \in \mathcal{L}$ 
(6)      retourner  $t$ 
(7)    pour tout  $T \in \mathcal{T}, t \in D_T$ 
(8)      empiler  $T(t), s(T(t)) = s(t) * S(T)$ 
(9)    fin
(10)  trier  $P$  par  $s$  décroissant
(11)  fin
(12)  retourner  $\lambda$ 
```

Entrée: $\mathcal{T}$ un ensemble de transformations, P une pile de candidats, d le nombre de dérivations maximum

Sortie: un analogue, ou rien

PCP-RECHERCHE-2($\mathcal{T}, P, d$)

```
(1)  si  $x \in \mathcal{L}$  alors retourner  $x$ 
(2)   $P \leftarrow x$  avec  $n(x) = 0$ 
(3)  tant que  $P \neq \emptyset$  et  $n(t) \leq d$ 
(4)    dépiler  $t$ , de note  $s(t), n(t)$ 
(5)    pour tout  $T, t \in D_T$ 
(6)      si  $T(t) \in \mathcal{L}$ 
(7)        retourner  $T(t)$ 
(8)      empiler  $T(t), s(T(t)) = E(T(t)), n(T(t)) = n(t) + 1$ 
(9)    fin
(10)  trier  $P$  par  $s$  et  $n$  décroissants
(11)  fin
(12)  retourner  $\lambda$ 
```

le cas de SR_2 , la construction des quatre fils de *lecteur* nécessite autant de recherches dans le lexique, auxquelles il faut rajouter les deux recherches correspondant à l'évaluation de *secteur*, soit au total, en comptant la recherche initiale, 7 appels à la fonction de recherche. Le nombre de nœuds dont on calcule la descendance est toutefois inférieur, puisque seuls deux nœuds ((0) et (1)) sont évalués. La sélection de l'un ou de l'autre critère de classement des candidats conditionne donc l'exploration de l'arbre, dans le sens d'une minimisation du nombre de recherches dans le cas SR_1 , et d'une minimisation du nombre de candidats engendrés dans le cas SR_2 .

Construction d'une prononciation

Une fois le ou les analogue(s) exhibé(s), ainsi que la série $T_{G_1} \dots T_{G_n}$ des transformations qui permettent de le(s) dériver du mot à transcrire, la construc-

SR_1 : chercher à bon escient							
lecteur	1	lection	0,5	secteur	0,05	section	0,025
		secteur	0,05	lecteuse	0,01	lecteuse	0,01
		lecteuse	0,01	flecteur	0,001	secteuse	0,005
		flecteur	0,001			flecteur	0,001
SR_2 : engendrer à bon escient							
lecteur	1	secteur	0,51	section	1		
		flecteur	0,51	flecteur	0,51		
		lecteuse	0,01	lecteusé	0,05		
		lection	0	secteusé	0,05		
				lection	0		

TAB. 4.7 – À la recherche de section.

tion de la prononciation s'exécute en appliquant en série les inverses des transformations phonétiques couplées aux T_{G_i} à la prononciation du ou des analogue(s), en respectant le principe énoncé à la section 4.4.2 : pour une transformation graphique donnée, les transformations phonétiques sont appliquées par taille de contexte décroissant. En d'autres termes, si T_{P_1} et T_{P_2} vérifient $D_{T_{P_2}} \subset D_{T_{P_1}}$ et $D_{T_{P_2}} \neq D_{T_{P_1}}$, alors T_{P_1} ne s'appliquera qu'aux formes auxquelles T_{P_2} ne s'applique pas. Plusieurs cas de figure sont alors possibles :

- La prononciation du mot inconnu est unique, et c'est elle qui est proposée.
- Compte-tenu de la multiplicité des transformations phonétiques associées à une transformation graphémique, plusieurs solutions sont possibles. Ces solutions sont alors classées en raison du produit des notes des transformations phonétiques qui ont contribué à leur élaboration.
- Aucune transformation ne peut être produite. Ceci se produit lorsqu'aucune des transcriptions élaborées par application des T_{P_i} n'appartient au domaine de définition d'une quelconque des $T_{P_{i+1}}$. Dans ces conditions, l'analogue est abandonné, et la recherche, reprise.

Illustrons le fonctionnement de ce mécanisme en nous appuyant sur l'exemple de la transformation $x\text{-}ly \rightarrow x$ dont les corrélats phonétiques sont listés dans la table 4.3 (en page 172). Supposons que le mot inconnu soit *worldly*, et que l'analogue exhibé au terme de la recherche soit *world*, transcrit /wɜ:ld/. Les transformations phonétiques applicables à cette forme sont celles notées T_{P_1} et T_{P_5} dans la table sus-mentionnée. L'application de ces transformations conduit à la construction de deux prononciations possibles pour *worldly* : /wɜ:ldli/ et /wɜ:ldi/. Nous sommes donc dans un cas de figure où plusieurs transcriptions sont en concurrence : compte-tenu des notes respectives des deux transformations qui conduisent à ces solutions, c'est la première transcription qui est (correctement) sélectionnée.

L'algorithme de prononciation

Maintenant que les différents composants de notre programme ont été décortiqués, il est possible de donner finalement une présentation synthétique de

4.4 Prononcer par application de transductions paradigmatiques 185

notre programme de transcription par cheminement paradigmatique : c'est le propos de l'algorithme 7.

Algorithme 7: Prononciation par cheminement paradigmatique

Entrée: un lexique $\mathcal{L}$ d'exemples, une chaîne x à transcrire

Sortie: une transcription pour x

```
PCP( $\mathcal{L}, x$ )
(1)  $\mathcal{T} \leftarrow \emptyset$ 
(2) pour tout  $y \in \mathcal{L}$ 
(3)    $\mathcal{T} \leftarrow \mathcal{T} \cup \text{PPC-EXTRACTION}(y)$ 
(4) fin
(5)  $\text{PCP-EVALUATION}(\mathcal{T})$ 
(6)  $P \leftarrow x$  avec  $n(x) = 0$ 
(7)  $\phi(x) \leftarrow \lambda$ 
(8) tant que  $\phi(x) = \lambda$ 
(9)    $z \leftarrow \text{PCP-RECHERCHE}(\mathcal{T}, P)$ 
(10)  si  $z \neq \lambda$ 
(11)    $\phi(x) \leftarrow \text{PCP-PRONONCE}(\mathcal{T}, z)$ 
(12) fin
(13) retourner  $\phi(x)$ 
```

4.4.3 Résultats préliminaires

Malgré nos efforts, et l'utilisation de conditions d'arrêts relativement draco-niennes, le principe de cheminement paradigmatique tel que nous l'avons im-planté dans PCP s'est révélé particulièrement gourmand en temps de calcul, spécialement lorsque la stratégie de recherche SR_2 est utilisée. La faute en in-combe principalement aux mots importés, pour lesquels, nous le verrons à la par la suite, la recherche d'un analogue est inéluctablement vouée à l'échec. Il faut hélas, pour s'en apercevoir, examiner exhaustivement la liste des candidats-analogues qui sont atteignables pour une condition d'arrêt fixée, ce qui peut se révéler très coûteux. En conséquence, les résultats que nous présentons dans cette section ne concernent que les seuls lexiques NETTALK, BDLEX, ainsi que le lexique ENST pour une évaluation sur un très gros lexique.

Résultats globaux

Les performances globales obtenues en utilisant PCP sont listées dans le ta-bleau 4.8. Ces résultats correspondent à la mise en œuvre des stratégies d'éva-luation SR_1 (avec une probabilité seuil de 10^{-2}) et SR_2 (avec une profondeur de recherche de 2).

Un premier résultat important est la diminution très sensible des taux de silence, par rapport aux résultats obtenus avec PPP : pour les deux lexiques, la stratégie SR_2 permet de diviser par deux le pourcentage de mots non-transcrits. Ce résultat remarquable est obtenu avec des contraintes d'exploration très fortes, trop forte probablement pour que la stratégie SR_1 , avec laquelle on obtient des taux de silence *supérieurs* à ceux obtenus avec PPP, puisse être convenable-ment évaluée. Le taux de silence important observé sur le lexiques BDLEX est dû

Corpus	Par mot			Par symbole	
	Correction	Silence	Justesse	Correction	Justesse
NETTALK (SR_1)	38,038	41,683	65,248	47,183	93,012
NETTALK (SR_2)	54,486	14,800	63,953	76,115	93,117
BDLEX (SR_1)	35,230	57,834	83,533	35,420	97,748
BDLEX (SR_2)	64,722	19,712	80,615	72,721	96,478

TAB. 4.8 – Performances de PCP sur différents lexiques.

pour une bonne part aux nombreux mots composés que contient ce lexique, et qui comptent pour environ un tiers du silence. Si l'on ne tient compte que des mots simples, le taux de silence est ramené en dessous de 15 %, ce qui est plus raisonnable. Un croisement des résultats des deux stratégies paradigmatiques permet de contrôler qu'environ 60 % des mots non-transcrits par PPP sont prononcés dans PCP; qui se répartissent approximativement en deux tiers d'erreurs, et un tiers de succès.

L'autre résultat important porte sur les performances en termes de justesse. Pour les deux lexiques, on retrouve des performances très proches de celles obtenues en utilisant des techniques de recollement de segments (voir section 3.3.3), soit des scores sensiblement inférieurs à ceux obtenus en utilisant PPP. La différence principale, c'est que ces nouveaux résultats ont été obtenus sans qu'il soit nécessaire d'aligner au préalable les formes graphémiques et phonétiques. L'examen plus détaillé des erreurs nous suggère par ailleurs que des améliorations simples sont susceptibles de faire disparaître un bon nombre d'erreurs. Ces résultats sont donc dans l'ensemble encourageants, et témoignent de la pertinence de l'approche par analogie.

Évaluation à la limite Nous avons fait figurer, sur le graphique 4.12, la courbe d'apprentissage de PCP évaluée sur le lexique de patronymes. Cette courbe a été obtenue en appliquant l'algorithme complet (extraction et restitution) jusqu'à une taille de lexique égale à 45 000 éléments. Au delà, le jeu de transformations a été gelé, et utilisé pour toutes les expériences impliquant des lexiques de taille supérieure. Pour ces expériences, la stratégie de recherche SR_2 a été uniformément utilisée, avec une profondeur de recherche de taille 2; on a par ailleurs filtré toutes les transformations graphiques attestées par moins de 10 paires.

La comparaison de cette courbe d'apprentissage avec celle de PPP est révélatrice de ce que cette généralisation du principe paradigmatique apporte. En particulier elle met clairement en évidence l'accélération de la réduction du taux de silence: là où il fallait plus de 50 000 exemples pour descendre sous les 10 % de silence, il faut, avec PCP, moins de 30 000 exemples. À la limite, on obtient une correction et une justesse légèrement inférieures à celle de PPP (respectivement 81,59 contre 85,33 et 87,96 contre 90,45). Notons que ce dernier score de justesse est légèrement supérieur à celui de DEC (85,05) pour la même tâche. De surcroît, l'allure de la courbe de justesse présente une pente relativement marquée, y compris pour des lexiques très volumineux, signe que cette stratégie permet de tirer en permanence avantage des exemples additionnels qui viennent

4.4 Prononcer par application de transductions paradigmatiques 187

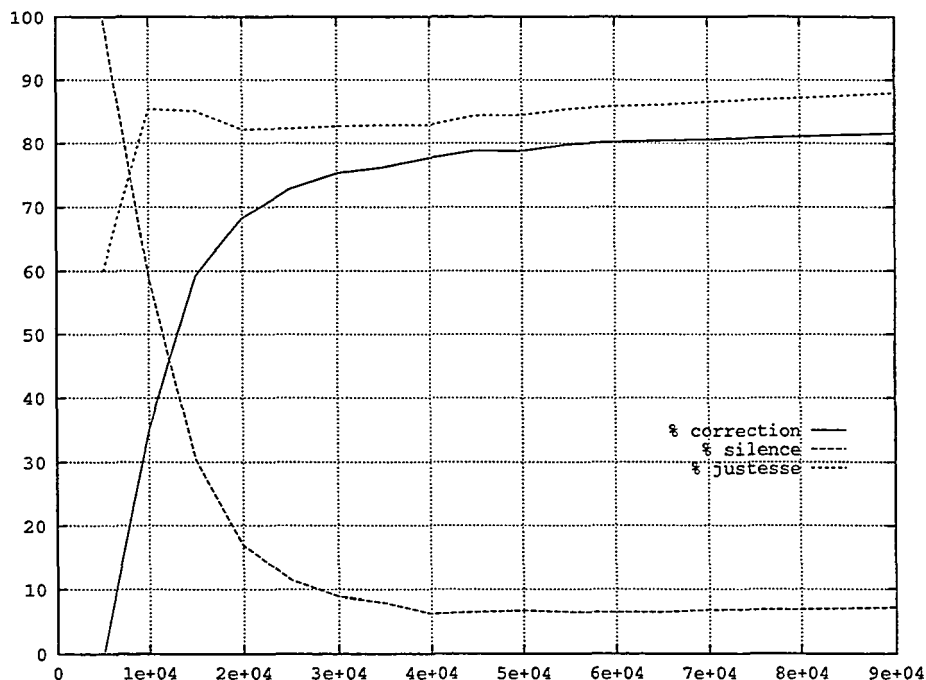


FIG. 4.12 – Évaluation de la vitesse d'apprentissage de PCP sur le lexique ENST.

enrichir la base d'exemples.

L'influence de quelques paramètres

Nous donnons dans cette section quelques indications concernant l'influence de quelques paramètres importants de PCP. Dans la mesure où l'exploration exhaustive de l'espace des paramètres n'a pas été entreprise, nous nous contentons, dans l'attente d'une étude plus fine, de quelques remarques qualitatives issues d'expérimentations sur des petits lexiques de tests (200 mots tirés aléatoirement dans le corpus NETTALK; l'ensemble d'apprentissage contient 18 000 éléments).

Conditions d'arrêt Indépendamment de la stratégie de recherche utilisée, les conditions d'arrêt fixent des seuils à la profondeur d'exploration de l'arbre des dérivés. Dans le cas de la stratégie SR_1 , ce seuil s'exprime en termes de la probabilité que le dérivé construit appartienne au lexique; dans le cas de SR_2 , ce seuil fixe une distance maximale en nombre de transformations entre un mot inconnu et un mot du lexique. Augmenter la profondeur de la recherche conduit, dans tous les cas, à diminuer le silence, puisque plus de solutions sont évaluées. Les gains que nous avons enregistrés sont toutefois faibles, et les solutions proposées tendent à être moins fiables au fur et à mesure que l'on empile le nombre de transformations à appliquer. Des évaluations complètes que nous avons menées, il ressort que le seuil utilisé pour la stratégie SR_1 est manifestement trop contraignant. Des expériences complémentaires ont montré qu'une augmentation de la

profondeur de l'exploration, réalisée conjointement à un filtrage fréquentiel plus brutal des transformations, conduisait à des taux de silence plus raisonnables. Ceci reste toutefois à confirmer sur des lexiques de tests plus conséquents.

Notons que dans les deux cas, la fixation de seuils aussi contraignants nous a *de facto* interdit de pouvoir transcrire des dérivés complexes, tels que *synergistically*, que l'on devrait pouvoir prononcer sur la base de *energy* par le chemin de transformations (toutes sont largement attestées dans le lexique NETTALK) : *synergistically* → *synergistical* → *synergistic* → *synergist* → *synergy* → *energy*.

Stratégie d'exploration Compte-tenu des conditions d'arrêt très restrictives que nous avons été amené à utiliser, les meilleurs résultats ont été obtenus grâce à l'utilisation la stratégie dite « engendrer à bon escient ». Une des raisons principales de succès de cette option est que, dans la mesure où la qualité des analogies exhibées laisse parfois à désirer, cette option qui conduit souvent à combiner l'influence de plusieurs analogues est moins dépendante du résultat de la recherche que la stratégie alternative, qui ne fonde la prononciation que sur un unique analogue.

L'augmentation de la qualité des transcriptions produites en fonction de la taille du groupe d'analogues est illustrée par le graphique de la figure 4.13, qui représente le pourcentage de mots correctement transcrits en fonction de la taille du voisinage.

Le taux de réussite croît manifestement en fonction du nombre d'analogues, et les performances les moins satisfaisantes sont obtenues lorsque la transcription ne se base que sur la transformation d'une seule entrée lexicale. Pour un lexique d'exemples de la taille de BDLEX, ce cas-là est malheureusement de loin le cas le plus fréquent, puisqu'il représente plus de 40 % des solutions produites.

Filtrage Le filtrage fréquentiel des transformations trop rares s'est avéré conduire à des résultats contrastés. Dans le cas de la stratégie de « rechercher à bon escient », le filtrage a toujours conduit à augmenter les performances en termes de justesse par mot; ce n'est pas ce que l'on a observé avec l'autre stratégie, où la réduction du nombre de transformations réduit mécaniquement la taille du groupe d'analogues mis à contribution, et entraîne de manière imprévue une moins grande justesse des transcriptions. Dans les deux cas, le taux de silence, lui, augmente.

Analogues et analogies

Un examen détaillé des mots qui n'ont pu être transcrits permet de retrouver une des propriétés une des propriétés intéressantes des approches analogiques déjà mise en valeur pour JUPA (voir section 3.2.3), à savoir leur capacité à identifier avec une certaine pertinence les mots lexicalement isolés. La table 4.9 reproduit la partie terminale de la liste des mots sans analogue, et donc impronçables, dans les expériences menées sur le lexique BDLEX.

Comme on peut le constater, cette liste contient de nombreux emprunts (*peeling*, *pronunciamento*) ou mots savants (*tmèse*, *isoglosse*), qu'il est, dans une

4.4 Prononcer par application de transductions paradigmatiques 189

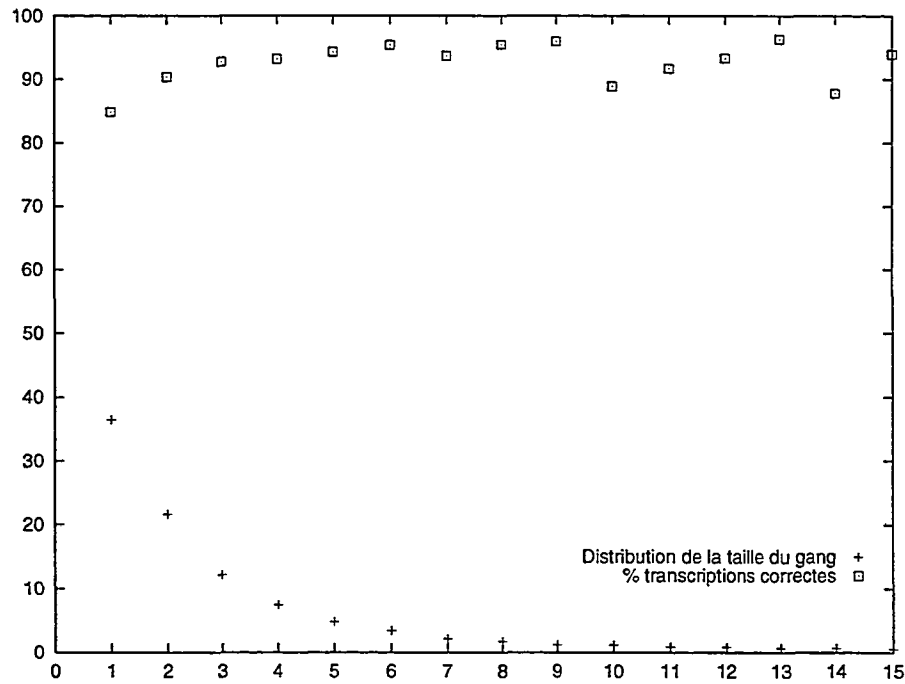


FIG. 4.13 – Pourcentage de succès en fonction du nombre d'analogues. Les données sont celles de BDLEX.

Le graphe est restreint aux mots qui ont au plus 15 analogues. Les effectifs observés au delà de cette valeur sont trop faibles pour être réellement significatifs.

engoulevant	flaccidité	footing
peeling	harcèlement	mineral
brièveté	isoglosse	clairsemé
ichtyosaure	djinn	dégingandé
tmèse	hamster	stockfisch
ellipsoïdal	pronunciamento	looping
arnica	puîné	ouvriérisme
western	inconcevable	ailloli
susmentionné	abscisse	manufacture
nénuphar	altercation	enorgueillir

TAB. 4.9 – Quelques mots sans analogue. Les extraits proviennent du lexique BDLEX.

Mot inconnu	Analogue	Mot inconnu	Analogue
toquade	toque	mascarade	mascaret
accréditer	créditer	âcre	âcreté
maroquinier	maroquin	revirement	virer
nativité	natif	pétulance	pétulant
suspecter	suspect	écho	échetier
iconographie	monographie	dénatalité	natalité
accentuer	accent	préscolaire	solaire
cosinus	minus	unisexe	unisexue
surélever	élever	septième	sept
synchrone	synchroniser	blèvement	bléser
innocence	innocent	endenter	chandelle
tuberculeux	tubercule	octal	octante
pianoter	piano	rengrènement	engrènement
satiété	piété	corvette	corvée

TAB. 4.10 – Quelques analogies extraites par PCP. Les extraits proviennent du lexique BDLEX.

certaine mesure, souhaitable de ne pas transcrire. On notera que lorsque des analogues acceptables sont disponibles, des mots d'origine étrangère peuvent être prononcés : ainsi *dumping* est prononcé par analogie avec *jumping*. La présence également dans cette liste de termes tels que *brièveté* (dérivé de *bref*) ou *flaccidité* (dérivé de *flasque*), qui s'inscrivent à la marge de paradigmes dérivationnels réguliers, nous montre les limites de l'approche paradigmatique pour une langue telle que le français, où les schémas dérivationnels intègrent fréquemment dans leur forme graphique des modifications phonologiques (c.f. *ouvriérisme*), et des scories étymologiques (*flasque* provient du latin *flaccidus*).

L'étude des mots pour lesquels un analogue *unique* a été exhibé, et dont quelques exemples choisis au hasard figurent dans la table 4.10 permet de mettre en évidence le relatif succès de nos procédures de recherche.

Pour tous les éléments de cette liste, en effet, chaque fois qu'un mot de la même famille est disponible dans le lexique, il a été utilisé par la procédure analogique : ainsi les paires (*accréditer*, *créditer*), (*nativité*, *natif*), associant des mots dont l'un dérive simplement de l'autre; ou encore (*revirement*, *virer*), qui fournit un exemple de dérivation plus complexe.

En l'absence de mot de la même famille, PCP se fonde sur des similarités plus hasardeuses de rime, ainsi les paires (*satiété*, *piété*) et (*cosinus*, *minus*), plus rarement d'attaque (*corvette*, *corvée*).

Typologie sommaire des erreurs

Une étude des erreurs commises sur le lexique BDLEX permet de les catégoriser en trois grandes familles. On retrouve tout d'abord les erreurs commises sur les mots irréguliers, qui sont dans tous les cas difficiles à éviter : ainsi *fart* transcrit /far/, *suber* transcrit /sybe/... Cette classe d'erreur irréductible est toutefois relativement restreinte comparativement aux autres.

4.4 Prononcer par application de transductions paradigmatiques 191

On trouve ensuite les mots réguliers dont la prononciation est construite par analogie avec un mot irrégulier : ainsi *nord*, transcrit /nɔʁd/ par similarité avec *lord*, ou *gambade* prononcé /gãmbad/ sur la base de *samba*. Ces erreurs-ci sont également relativement rares.

L'écrasante majorité des erreurs provient en fait de la trop grande simplicité du mécanisme de transcription; en particulier l'utilisation de règles indépendantes du contexte conduit à passer à côté d'un bon nombre de transformations morphologiques régulières qui sont phonologiquement conditionnées. Un exemple caractéristique de ce type d'erreurs est fourni par la préfixation de *in* ou *im*: la transformation $X \rightarrow in-X$ a deux corrélats phonétiques majeurs : $/X/ \rightarrow /ẽ-X/$ et $/X/ \rightarrow /in-X/$, selon que *X* débute par une consonne non-nasale ou par une voyelle. Comme ces deux transformations phonétiques ont le même domaine de définition (non-contraint), la première, qui est majoritaire s'applique uniformément, conduisant à des formes telles que /ẽegal/ (pour *inégal*), /ẽnomabl/ (pour *innomable*),... Le même phénomène est à l'origine de très nombreuses confusions sur les mots qui comportent ou non des *h* aspirés; de l'incohérence des distributions entre voyelle et semi-voyelle pour les voyelles hautes; de nombreuses erreurs sur l'interprétation de *c* (/s/ ou /k/) et de *g* (/g/ ou /ʒ/)... Une des priorités pour améliorer de manière certaine et sensible les résultats est donc de modifier cet aspect de la procédure. Compte-tenu de la multiplicité des techniques disponibles pour extraire des règles dépendant du contexte, il nous semble qu'une telle amélioration ne pose pas de réelle difficulté technique.

4.4.4 Conclusion et perspectives

Nous avons présenté dans cette section un algorithme original fondé linguistiquement sur le concept de paradigme, et qui fonctionne sur le principe du couplage entre transformations internes à deux domaines distincts, le domaine graphique d'une part, le domaine phonétique d'autre part. Nous avons appliqué cet algorithme au problème de la transcription graphème-phonème, et présenté un certain nombre d'éléments d'évaluation. Ces premiers résultats nous ont permis de montrer la possibilité algorithmique et l'efficacité d'une telle approche.

En particulier, la démarche mise en œuvre par l'algorithme PCP nous semble présenter un certain nombre de caractéristiques satisfaisantes :

- elle permet de s'affranchir de l'étape d'alignement préalable des données. Nous avons montré (c.f. la section 2.4.4) que ce pré-traitement, indispensable à la mise en œuvre de techniques classificatoires, fournit au système d'apprentissage un biais important, et restreint le champ d'application de ces techniques à des transductions entre chaînes suffisamment « ressemblantes » pour qu'une correspondance terme à terme puisse être exhibée. Notre mécanisme de transcription permet de se dispenser de cette étape, ce qui le rend en particulier susceptible de s'appliquer à d'autres tâches exprimables dans les termes d'une transduction. Pour illustrer ce dernier point, nous avons réalisé quelques expériences sur la tâche de transcription phonème-graphème, et obtenus, sur le lexique NETTALK, des scores de justesse par lettre supérieurs à 92 %, correspondant à une justesse par

mot supérieure à 58 %. Une autre utilisation de cette caractéristique de notre algorithme serait de manipuler non plus des chaînes de phonèmes, mais d'intégrer de manière relativement naturelle des informations telles que la position de l'accent de mot. Ceci n'est possible, dans une optique classificatoire, qu'au prix du doublement du nombre des voyelles reconnues (une version accentuée et une version non-accentuée), ce qui contribue à augmenter le nombre de classes à prédire, et à dégrader les performances;

- elle parvient à isoler de manière relativement pertinente les formes graphiques « atypiques », notamment les emprunts, par rapport au lexique d'exemples qui est disponible. Pour autant, elle semble à même d'extraire un certain nombre de « micro-régularités » (c.f. la transcription de *jumping*), dès lors que ces régularités présentent un mode d'organisation paradigmatique;
- elle autorise la prise en compte « naturelle » d'informations supplémentaires, notamment de celles qui s'expriment au niveau d'une graphie. En particulier, la prise en compte des catégories morpho-syntaxiques, qu'il est possible d'utiliser pour typer les transformations extraites, semble offrir des perspectives d'amélioration intéressantes;
- elle conduit à des performances de transcription qui sont, en termes de justesse, très proches de celles obtenues avec SMPA sur le lexique NETTALK, et bien supérieures à celles obtenues avec DEC avec ces mêmes données. Ces résultats sont d'autant plus encourageants qu'ils sont obtenus à partir de données non-alignées. Les courbes d'apprentissages observées sur des gros lexiques témoignent par ailleurs une capacité à mieux rentabiliser les exemples supplémentaires fournis au système;
- elle fournit un modèle de prononciation par analogie sensiblement plus satisfaisant que ceux qui ont pu être proposés dans le passé, dans la mesure où elle propose une caractérisation formelle de la notion d'analogie. Par ailleurs, elle permet de modéliser un certain nombre d'effets observés chez l'humain, tel que l'effet de « gang ». À l'appui de cette affirmation, on peut mentionner les résultats d'une petite expérience conduite avec les 70 pseudo-mots de Glushko. Nous avons transcrit ces pseudo-mots, qui sont donnés en annexe de (Dedina & Nusbaum 1991), avec l'algorithme PCP, en utilisant les données complètes de NETTALK comme données d'apprentissage. Sur ces 70 pseudo-mots, 69 sont transcrits par PCP, parmi lesquels 9¹² seulement de manière incorrecte (i.e. différemment des prononciations attestées lors d'expériences de lecture), soit un taux de réussite de 87 %. Si ces résultats sont légèrement inférieurs à ceux publiés par Dedina et Nusbaum, ils n'en restent pas moins encourageants, compte-tenu d'une part de l'absence d'alignement, et d'autre part de la différence entre les bases de données utilisées.

Des extensions de ce modèle sont possibles, comme par exemple celle qui consisterait à intégrer les fréquences des lexies dans la note affectée à des

12. Trois erreurs sont d'ailleurs assez discutables: *bild* et *pild*, que notre système prononce sur le modèle de *wild*, alors que seuls */bild/* et */pild/* sont attestés, et *pomb*, que notre système prononce par analogie avec *tomb* (*/tu:m/*), alors que seuls */pəʊm/* et */paɪm/* sont observés.

transcriptions concurrentes. De telles extensions seraient, nous semble-t-il, nécessaires pour préciser la pertinence d'un tel modèle dans un cadre cognitif.

En résumé, la démarche analogique mise en œuvre dans notre modèle répond à bon nombre des critiques que nous adressions à la section 2.4 aux méthodes classificatoires, et, inversement, semble atteindre les objectifs que nous imposions à un système de transcription par analogie au commencement de cette étude. Seule tâche d'ombre, les performances en termes de correction sont encore, à nos yeux, insuffisantes. L'examen des erreurs les plus manifestes commises par notre algorithme nous a néanmoins convaincu que des améliorations simples (typage morpho-syntaxique des transformations graphiques; extraction des transformations phonétiques sous la forme de règles dépendant du contexte; utilisation de modèles de langage ou de grammaires phonotactiques pour affiner l'ordre d'évaluation des candidats; prise en compte simultanée de plusieurs solutions dans le cadre de la procédure de recherche SR_1) étaient envisageables, qui entraîneraient probablement une augmentation sensible des performances, sans qu'il soit nécessaire de remettre en cause le type de représentation des données, et en particulier de revenir sur notre décision de ne pas utiliser, pour l'apprentissage, de transcriptions alignées.

Nous aurons l'occasion de revenir plus longuement sur les perspectives générales qu'ouvre cette approche, en la resituant dans le contexte général des techniques d'apprentissage pour le traitement du langage naturel à la section 4.6.

4.5 La conjugaison des verbes anglais

4.5.1 Présentation du problème

Introduction

Dans le but de préciser la validité de l'approche fondée sur l'extraction et l'exploitation de relations paradigmatiques, telle qu'elle est mise en œuvre dans PCP, nous avons souhaité l'appliquer à un autre type de problème, pour lequel cette démarche nous semblait également pertinente, à savoir la prédiction, à partir d'une base d'exemples, des formes passées des verbes anglais.

Cette tâche présente toutes les caractéristiques inhérentes au traitement du langage naturel, caractéristiques dont nous avons vu qu'elles faisaient tout le sel de l'apprentissage d'un système de transcription graphème-phonème. En effet, la description linguistique de ce problème mobilise traditionnellement, comme pour le problème de la transcription, les éléments suivants: un petit ensemble de règles contextuelles (ici des règles de flexion) massivement attestées, qu'il est toutefois nécessaire d'étendre pour prendre en considération l'existence de nids d'exceptions présentant localement la forme de micro-paradigmes, et un petit nombre d'idiosyncrasies.

Le problème de l'apprentissage du système de conjugaison de l'anglais présente, en sus de son intérêt intrinsèque, l'avantage d'avoir été largement uti-

lisé comme banc d'essai pour différents modèles ou algorithmes d'apprentissage. Depuis l'implémentation connexionniste d'une architecture dédiée à cette tâche (Rumelhart & McClelland 1986), abondamment discutée et critiquée dans (Pinker & Prince 1988) et dans (Lachter & Bever 1988), diverses approches alternatives ont été proposées : une nouvelle implantation connexionniste, sans doute moins discutable que celle de Rumelhart et McClelland, est décrite dans (MacWhinney & Leinbach 1991); une approche symbolique fondée sur l'utilisation d'arbres de décision est exposée dans (Ling & Marinov 1993; Ling 1994), et fournit des éléments précis de comparaison; et plus récemment quelques résultats ont été obtenus en utilisant des techniques de Programmation Logique Inductive (ILP) (Califf & Mooney 1995). On notera également que le même problème est traité dans un modèle à base d'analogie¹³ dans (Skousen 1989), qui s'intéresse à la formation des formes passées des verbes finlandais, dont la morphologie est notoirement complexe. En résumé, les techniques d'apprentissage auxquelles ce problème a été soumis sont sensiblement les mêmes techniques que celles qui ont été passées en revue dans le cadre de la transcription graphème-phonème au chapitre 2.

Ces deux caractéristiques de la tâche de conjugaison des verbes anglais contribuent à en faire un banc d'essai tout désigné pour notre algorithme analogique.

Dans cette section, après avoir rappelé les difficultés posées par l'induction de la forme passée d'un verbe anglais à partir de sa forme radicale (première personne du présent), nous présentons et commentons les résultats correspondant à différentes mises en œuvre de l'approche analogique pour traiter ce problème particulier. Notre ambition est triple, puisque nous voulons d'une part prouver que l'approche analogique est pertinente pour d'autres problèmes d'apprentissage du langage naturel que celui de la transcription graphème-phonème; d'autre part étudier plus finement, grâce aux données très complètes dont nous disposons (voir l'annexe I), les forces et faiblesses de l'approche analogique; et enfin vérifier pour vérifier le bien fondé de l'intuition que nos algorithmes sont des réponses adaptées à la tâche d'*analyse* morphologique de formes inconnues.

En revanche, nous ne nous sommes pas intéressé ici à la question de savoir si le type d'acquisition de connaissances mis en œuvre dans PCP était ou non un modèle cognitivement plausible de l'apprentissage par l'enfant des schémas inflexionnels; question qui a été largement discutée à propos des modèles d'apprentissages évoqués ci-dessus. Nous nous dispensons donc de la tâche qui consisterait à élaborer un protocole d'apprentissage plausible, qui, par le biais de l'utilisation des fréquences relatives des différents verbes, permettrait de simuler de façon réaliste la régulière et progressive exposition de l'enfant à des formes verbales de plus en plus rares, et de reproduire des phénomènes tels que la très controversée « courbe en U » (Lachter & Bever 1988; Plunkett & Marchman 1990).

13. Voir toutefois la section 2.4.6 pour une discussion sur la conception de l'analogie, profondément divergente de la nôtre, que ce système met en œuvre.

La formation de la forme passée des verbes anglais

Pour cette description de la formation du passé, nous suivons pour l'essentiel la présentation donnée dans (Pinker & Prince 1988, p82-87), à une nuance près. La notion de *régularité*, qui tient une place importante dans leur description, est nous semble-t-il, en morphologie du moins, une notion toute relative. Il existe en effet bien des cas dans lesquels identifier le régulier de l'irrégulier est largement problématique (nous reprenons ici l'analyse de MacWhinney et Leinbach (1991, p137-138)). Il serait à nos yeux plus correct de ne pas utiliser ce terme, et de prendre acte du fait que de nombreuses langues attestent un état de concurrence entre plusieurs paradigmes flexionnels, parmi lesquels certains sont parfois plus productifs que d'autres. Ainsi le français identifie trois *groupes* de verbes, et le paradigme de conjugaison correspondant au premier groupe est sans nul doute le plus productif, comme en témoigne son fort pouvoir attracteur auprès des nouveaux (emprunts ou dérivés) verbes. Il nous semble toutefois difficile d'affirmer qu'en conséquence, les verbes du second groupe sont irréguliers. Dans le cas précis qui nous concerne, le déséquilibre entre les différents paradigmes est néanmoins tel que nous avons choisi, pour alléger la rédaction, de conserver ce terme.

Verbes Réguliers En anglais, la grande majorité des verbes construit son passé par suffixation de /d/, ce qui se traduit en surface de diverses manières, suivant la nature de la terminaison de la forme radicale. On observe en effet les trois réalisations suivantes :

- /ɪd/ à la finale des verbes finissant par une occlusive dentale, tels que *need*. La seule suffixation de /d/ conduirait pour ces verbes à la constitution de groupes /td/ ou /dd/, qui sont prohibés en anglais. L'insertion d'un /ɪ/¹⁴ entre les deux consonnes permet de restaurer l'acceptabilité du composé. On notera que, *mutatis mutandis*, le même mécanisme phonologique est également à l'œuvre dans la formation régulière de la troisième personne du singulier des verbes (ou du pluriel des noms) dont le radical se termine par une sibilante. Ainsi *pass*, qui s'inflecte en /pasɪz/, pour éviter la formation d'un groupe consonantique /sɪz/. L'épenthèse de /ɪ/ n'est donc en rien un trait caractéristique du processus inflexionnel, mais plutôt la manifestation d'une contrainte phonotactique qui prohibe systématiquement la succession de deux consonnes trop « ressemblantes ».
- /t/ à la finale des verbes dont le radical se termine par une obstruante non-voisée, tel que *push*. Ici encore, une règle phonologique (voire phonétique) est à l'œuvre, qui conduit à l'assimilation par le /d/ de la surdité de la consonne finale du radicale. Cette règle également s'applique quasiment indistinctement à tous les groupes de consonnes en anglais, et empêche la formation de groupes composés d'une obstruante non-voisée suivie d'une obstruante voisée.

14. Nous avons trouvé, selon les sources, plusieurs transcriptions différentes pour cette voyelle épenthétique: /ɪ/, /ə/ et /ɪ/. Nous nous en tenons à celle utilisée dans nos données, à savoir /ɪ/.

- /d/ dans tous les autres cas, c'est-à-dire quand le verbe se termine par une consonne voisée autre que /d/, ou par une voyelle.

La manière traditionnelle (i.e. générativiste) de décrire ce processus est de poser une règle (morphologique) de construction du dérivée (la suffixation de /d/), dérivé qui est ensuite transformé au cours de son cheminement dans le module phonologique de l'anglais par des règles qui ne sont pas spécifiques à la formation du passé (épenthèse de /ɪ/ entre consonnes de même lieu d'articulation, dévoisement, etc.)

Verbes Irréguliers À côté de ces formes régulières, un certain nombre de verbes (les verbes *forts*) construisent leur forme passée par diverses autres transformations morpho-phonologiques, qui se manifestent en surface de multiples manières. Sans rentrer de le détail de la typologie de ces verbes (voir (Pinker & Prince 1988, p185-187)), citons les principaux mécanismes mis en œuvre dans cette transformation :

- affixation nulle, correspondant à l'utilisation de la forme radicale y compris au passé. C'est le cas de verbes tels que *cut* ou *let*. Tous ces verbes ont une forme radicale se finissant en /t/ ou /d/;
- réalisation d'un /t/ final, résultat soit d'une pure suffixation (ainsi *burn*), soit du dévoisement d'un /d/ final (comme dans *send*);
- changement de la voyelle ou de la diphtongue du radical, que ce changement s'analyse comme une alternance voyelle tendue v.s. voyelle lâche (comme dans *meet* (/mit/ v.s. /mɛt/); ou non (phénomène d'*ablaut*) comme dans *speak* (/spik/ v.s. /spok/);
- application simultanée des deux transformations précédentes, à savoir suffixation d'un /t/ et changement de voyelle : c'est le cas de verbes tels que *mean* ou *buy*.
- supplétion complète, la forme passée est en surface sans aucune ressemblance avec la forme radicale. C'est le cas de *go*.

Remarques complémentaires Il est important de garder à l'esprit les quelques remarques suivantes, qui complètent la description du phénomène. Ces formes passées de l'anglais correspondent à quatre usages syntaxiques différents du verbe, qui sont le passé simple, le participe passé, la forme passive, et l'adjectif formé sur le verbe. Les verbes forts ne respectent pas complètement ce syncrétisme puisqu'un participe passé distinct existe pour de nombreux verbes (par exemple *speak*). Il existe par ailleurs pour les verbes forts un certain flottement concernant ces différents usages, avec une tendance à la régularisation : certaines formes (ou certains usages de certaines formes) irrégulières sont aujourd'hui considérées comme désuètes, et n'apparaissent plus que dans des contextes restreints (textes littéraires, etc). C'est par exemple le cas de *learn*, dont la forme irrégulière *learnt* tend à être supplantée par *learned*.

La résistance des verbes forts à la régularisation est une conséquence probable de la haute fréquence d'un bon nombre de ces verbes. Pour les verbes

moins fréquents, ils semblent résister d'autant mieux qu'ils présentent des analogies de surface avec des verbes irréguliers et fréquents (une rime commune, par exemple); ce qui leur permet de s'inscrire dans des « micro-paradigmes », tel que celui formé par les verbes dont le radical se termine par *ing*¹⁵. Enfin, un grand nombre de verbes forts partagent la caractéristique d'être mono-syllabiques.

Le type d'inflexion que suit un verbe est une propriété strictement lexicale de ce verbe. En particulier elle est conservée lors des dérivations morphologiques « directes », i.e. qui préservent la catégorie morpho-syntaxique du verbe, comme celle qui permet de dériver *mistake* de *take*. En particulier également, elle est perdue lors des transformations morphologiques qui impliquent un changement de catégorie morpho-syntaxique¹⁶: ainsi *grandstand* est régulier, puisqu'il dérive du nom *grandstand*, alors que *stand* (comme *understand*,...) est irrégulier. En conséquence, l'appartenance ou non d'un verbe à la catégorie des verbes forts ne peut se reconnaître à partir de l'observation de certaines de ses caractéristiques prises isolément, comme par exemple la réalisation phonétique (ou graphique) de son radical. Il existe en effet des verbes homophones (*wring/ring*), voire homographes et homophones (c.f. les deux usages de *hang*) qui néanmoins obéissent à des schémas inflexionnels différents.

Quelques chiffres Pour achever de poser le cadre du travail entrepris dans cette section, donnons les performances « brutes » obtenus pour cette tâche par diverses approches alternatives. Pour une partie, nous reprenons les résultats¹⁷ donnés dans (Ling & Marinov 1993, p. 270), bien que ces auteurs se placent dans des conditions expérimentales sensiblement plus confortables que nous (suppression des homophones et des verbes possédant plusieurs formes passées différentes, suppression des verbes de plus de 8 phonèmes), mais avec des ensembles d'apprentissage plus petits (500 verbes). Ces résultats, reproduits dans la table 4.11, ont été complétés par ceux obtenus par Mooney et Califf (1995). Quand cette distinction était disponible, nous avons séparé les performances obtenues sur les verbes réguliers (colonne **Rég.**) et irréguliers (colonne **Irr.**).

Même si, compte-tenu des motivations assez différentes qui ont conduits à ces résultats, ces chiffres ont peu de signification en eux-mêmes, ils permettent de donner une idée de la difficulté intrinsèque du problème.

15. Cette caractéristique de leur graphie ne suffit pas toutefois à identifier avec certitude leur irrégularité: certains verbes en *ing* sont réguliers, comme par exemple *wing*.

16. Plus précisément lors de dérivations qui entraînent la formation de mots dont la tête lexicale n'est plus le verbe (voir (Pinker 1994, p.144 et suivantes), pour de nombreuses illustrations de ce mécanisme, qui n'est pas limité aux dérivations verbales.)

17. Aussi bien pour les arbres de décision que pour les réseaux de neurones, des améliorations ont été suggérées pour améliorer sensiblement ces résultats: certaines configurations leur permettent ainsi de parvenir à un pourcentage de formes correctes égal respectivement à 80,0 % (réseau connexionniste) et 89,4 % (arbre de décision) (MacWhinney 1993). Dans la mesure où ces améliorations supposent que les données soient complétées par des informations supplémentaires (marquage des voyelles et des consonnes, représentation distribuée des phonèmes, etc.), que nous préférons pour l'instant ignorer, nous nous en tenons à ces résultats, qui, quoiqu'obtenus dans des conditions expérimentales sensiblement différentes des nôtres, nous semblent offrir les meilleures bases de comparaisons.

Méthode	Référence	Rég.	Irr.	Moyenne
Arbre de décision	(Ling & Marinov 1993)	84,9	10,0	76,2
Réseau connexionniste	(MacWhinney & Leinbach 1991)	37,3	4,1	33,4
Logique inductive	(Califf & Mooney 1995)	?	?	≈ 90,0

TAB. 4.11 – *État des performances sur l'apprentissage de la conjugaison de l'anglais*

4.5.2 Manipulations expérimentales et résultats

Dans les expériences que nous allons maintenant reporter, nous faisons abstraction des réserves que soulèvent la représentation simpliste des entrées lexicales par leur transcription phonétique. L'entrée de nos procédures, sauf mention explicite du contraire, est une chaîne phonétique représentant le verbe à conjuguer. Notons toutefois que la prise en compte de toutes les informations lexicales nécessaires à l'identification plus précise de la forme à transcrire ne pose aucun problème dans notre formalisme (c.f. la prise en compte des catégories morpho-syntaxiques reportée à la section 4.3.3).

Notons également que, à la suite des expériences de MacWhinney et Leinbach d'une part, et de Ling et Marinov d'autre part, nous avons également mesuré les performances de notre système face aux tâches de construction du passé composé d'une part, et de construction de la troisième personne du singulier d'autre part. Ces deux tâches présentent suffisamment d'analogies avec celle qui nous concerne en premier chef pour qu'il ne soit pas nécessaire de s'étendre plus longuement sur le type de difficultés qu'elles présentent.

Tous les résultats présentés ont été obtenus grâce à l'utilisation de l'algorithme PCP (voir section 4.4); dans tous les cas la stratégie de recherche utilisée a été celle décrite sous le nom de « engendrer à bon escient »; la distance maximale autorisée en termes de transformation a été fixée à 3; toutes les transformations sont mises à contribution durant la recherche (pas de filtrage fréquentiel). Une présentation complète des données utilisées est fournie en annexe I.

La flexion comme une transduction hétéro-alphabétique

Résultats globaux La manière la plus simple d'appliquer nos procédures au problème de la formation des verbes anglais est d'interpréter ce problème dans les mêmes termes que le problème de la conversion graphème-phonème, en faisant jouer aux formes radicales le rôle des formes graphiques, et aux formes conjuguées le rôle des formes phonétiques. Dans cette perspective, nos procédures fonctionnent « comme si » les entrées et les sorties du système appartenaient à des espaces de représentation différents, ce qui n'est naturellement pas le cas, puisque radicaux et forme conjuguées sont représentés par des chaînes phonétiques. La conjugaison est donc appréhendée comme un simple processus de transduction d'une chaîne de symboles d'un alphabet *A* dans une chaîne de symboles d'un alphabet *B*, les deux alphabets étant supposés, comme dans le cas de la transcription, totalement distincts.

Le type de transformations que nous pouvons extraire dans ce contexte re-

pose alors uniquement sur la ressemblance entre les formes radicales. Un exemple d'une telle transformation est transformation initiale /t/ → /b/, attestée par les formes phonétiques des paires alternantes *turn/burn*, *tear/bear*, etc.

Par contre, confronté à une telle présentation des données, PCP est incapable de réaliser que la construction du passé respecte un schéma flexionnel grossièrement paradigmatique. En conséquence, dans cette expérience, l'apprentissage du fait que la forme passée se construit (en gros) par suffixation de /d/ est impossible. Au contraire, pour un radical donné, la prédiction de la forme fléchie se fait par analogie avec d'autres formes radicales présentant des similarités (rime ou attaque de mot identique).

Le tableau 4.12 donne les résultats obtenus pour les trois tâches considérées. Dans tous les cas, les résultats sont des pourcentages moyens sur 10 tests. Dans le cas du passé simple et du passé composé, chaque expérience comprenait l'extraction de transformations à partir de 1000 exemples, 400 verbes étant mis de côté pour les tests. Les chiffres sont respectivement de 500 et de 160 pour la tâche de construction de la troisième personne du singulier, à cause du nombre plus restreint d'exemples disponibles.

Tâche	Correction	Silence	Justesse
Passé simple	57,70	32,35	85,27
Passé composé	45,21	45,39	82,80
3 ^e pers. sing.	13,88	85,75	97,34

TAB. 4.12 – Performances de PCP pour les trois tâches de conjugaison. Les alphabets codant formes radicales et fléchies sont considérés comme distincts.

Si l'on fait abstraction des taux de silence importants, qui s'expliquent par la relative petitesse des lexiques d'exemples¹⁸, les taux de justesse obtenus sont relativement satisfaisants, et très comparables aux résultats présentés dans (Ling & Marinov 1993). La même échelle de difficulté pour les trois tâches est également reproduite, la prédiction du passé simple se trouvant à un niveau intermédiaire entre celle de la 3^e personne du singulier, et celle du passé composé.

Examen des erreurs dans la construction du passé La distribution des erreurs selon le caractère régulier ou irrégulier de la forme à prédire permet de mettre à jour des différences de justesse très importantes. Pour les verbes réguliers (2733 effectivement conjugués dans notre série de tests), le taux de justesse de PCP est en effet de 93,81 %, alors que pour les verbes réguliers (329 occurrences conjuguées), il est de 23,71 %.

Si l'on se penche sur les erreurs commises sur les verbes irréguliers, on remarque que la très grande majorité des erreurs commises par le programme sont des erreurs de régularisation. Dans tous les cas, toutefois, les verbes sont *correctement* régularisés, c'est-à-dire que les formes produites respectent systématiquement les règles phonologiques énoncées dans la section introductive. La petite minorité des autres erreurs correspondent à la production de formes

18. Les transformations que nous avons extraites dans ce contexte pourraient d'ailleurs l'être de n'importe quel lexique, et utilisables pour toute autre sorte tâche justiciable de la même stratégie analogique.

irrégulières tout à fait plausibles : ainsi *chose* par exemple est conjugué /tʃɔst/ sur le modèle de *lose*; *bring* est conjugué /bræŋ/ sur le modèle de *ring*; *take* est conjugué /tɔk/ sur le modèle de *wake*; et conséquence directe de notre seule connaissance des formes phonétiques, *write* est conjugué /rɔt/, par analogie avec *fight*...

Réciproquement, l'écrasante majorité des erreurs commises sur les verbes réguliers proviennent de leur inclusion erronée dans des sous-classes relativement productives de verbes irréguliers. Ceci est particulièrement vrai pour les verbes se terminant par une occlusive dentale, comme *need* conjugué comme *feed*, *blend* rapproché de *bend*... Il est important de noter que l'on a relevé dans cette classe d'erreurs une proportion infime (4 sur nos 47 erreurs, soit au total 4 sur 2706 formes conjuguées produites) de formes totalement invraisemblables, telles que /sendaist/ pour *sense*. Ces formes ont été produites par application en chaîne de plusieurs transformations, dont certaines correspondent probablement à des alternances incorrectes extraites à partir de paires verbe faible/verbe fort.

Le petit quart de verbes irréguliers pour lequel le mécanisme analogique a permis de retrouver la bonne conjugaison témoigne de la capacité de notre algorithme à tirer profit d'effets de « gang » pour les paradigmes minoritaires très homogènes phonétiquement (les verbes se finissant en /ɪŋ/ par exemple).

Un dernier point mérite d'être mentionné : s'il est vrai que la régularité ou non des verbes est imprévisible par simple examen de la forme phonétique du radical, la stratégie analogique que nous avons adoptée présente toutefois un mérite supplémentaire par rapport aux autres approches. En effet, dans la mesure où les transformations les mieux attestées sur notre lexique correspondent à des changements simples (exemplairement le remplacement d'une consonne initiale par une autre), les verbes pour lesquels PCP tend à produire une forme irrégulière sont en grande majorité des verbes monosyllabiques. D'une certaine manière donc, cette caractéristique importante des verbes irréguliers est correctement identifiée, et émerge comme un produit dérivé désirable dans nos résultats.

En conclusion, la stratégie strictement analogique que nous avons mise en œuvre fonctionne, lorsqu'elle fonctionne, comme si elle avait permis d'apprendre correctement, compte-tenu des imperfections des représentations utilisées, la morpho-phonologie de la formation du passé en anglais : les formes régulières produites respectent toutes les contraintes phonotactiques de surface qui obscurcissent la simplicité de la conjugaison dans le paradigme le plus productif; et un certain nombre de formes irrégulières sont (correctement ou incorrectement) prédites, qui d'une part respectent quasiment toujours des schémas flexionnels attestés, et d'autre part sont préférentiellement produites pour les verbes monosyllabiques.

La flexion comme une transduction homo-alphabétique

Réorganisation des exemples Nous avons souhaité dans cette section envisager une mise en œuvre alternative de l'approche analogique, qui permette à notre algorithme de prendre en compte le fait que les formes radicales et passées appartiennent à un même espace de représentation, celui des chaînes

phonétiques. Nous espérons par ce biais expliciter la correspondance qui existe entre radical et forme conjuguée, et par ce biais réduire le silence. Compte-tenu de la simplicité de la transformation impliquée dans le paradigme le plus productif, ce taux de silence est largement injustifié, dès lors que le mécanisme apprenant parvient à réaliser que tout ce qu'il suffit de faire, c'est essentiellement de rajouter un /d/ à la fin du radical à conjuguer.

En conséquence, nous avons dans cette expérience modifié le format des entrées à conjuguer, en introduisant des formes abstraites de la manière suivante. Pour chaque couple (forme radicale, forme fléchie) des lexiques d'apprentissage, nous avons créé deux exemples distincts du type (forme radicale abstraite, forme radicale) et (forme fléchie abstraite, forme fléchie). La forme abstraite est construite en ajoutant à la forme radicale un suffixe abstrait, marquant la conjugaison représentée par la forme fléchie. Ainsi le couple (/spɪk/, /spɒk/) contribue aux deux exemples suivants: (spɪk+R, /spɪk/) et (spɪk+D, /spɒk/), où +R et +D signalent respectivement les formes radicales et passées du verbe. Les formes radicales de tous les verbes figurent dans le lexique d'apprentissage.

On pourra nous objecter que cette manière de procéder introduit un biais manifeste, puisque le choix de *suffixer* les marqueurs abstraits signale que cette conjugaison se traduit au niveau phonologique par une suffixation. Une alternative aurait été d'utiliser en entrée les formes graphiques, laissant à l'algorithme le travail d'identifier le lieu de l'affixation. Compte-tenu toutefois de l'irrégularité de la formation des formes passées correspondant aux graphies (changement de y en ied, doublement de certaines consonnes), nous avons considéré que cette manière de procéder conduirait à l'introduction d'un bruit supplémentaire dans les données, et nuirait par là-même à notre démonstration.

En procédant de cette façon, PCP est à même d'identifier, en plus des transformations correspondant à des paires minimales dans l'espace des formes radicales, une transformation supplémentaire, massivement attestée: $+R \rightarrow +D$, correspondant aux couples (forme radicale abstraite, forme fléchie abstraite). La recherche de l'analogie est alors le fruit d'une compétition entre la transformation réellement morphologique (qui est toujours couronnée de succès, puisque le radical est toujours disponible), et des transformations strictement analogiques. Ce que nous espérons mieux capturer ainsi, c'est l'effet de concurrence entre la pression analogique exercée par des verbes « hypersimilaires » (c.f. (Pinker & Prince 1988, p116)), qui s'oppose à l'application du paradigme flexionnel dominant.

Cette manière de procéder présente un avantage supplémentaire, puisqu'elle permet de traiter simultanément toutes les conjugaisons, en conservant l'indication que différentes formes peuvent appartenir à un même verbe. Pour cette expérience-là, dont les résultats figurent à la ligne « conjugaison complète » dans le tableau 4.13, nous avons aléatoirement séparé l'ensemble de toutes les formes abstraites (y compris celles correspondant au radical et au gérondif) en deux groupes, un pour l'apprentissage et un pour le test. On notera que dans ce test, une entrée inconnue peut correspondre à la forme radicale d'un verbe.

Nous avons finalement testé la capacité de PCP à effectuer une tâche d'analyse, en inversant entrées et sorties de la tâche de conjugaison complète. Dans ce test, dont le résultat figure à la ligne « Analyse complète » du tableau 4.13,

Tâche	Correction	Silence	Justesse
Passé simple	86,350	0,425	86,718
Passé composé	83,070	1,163	84,043
3 ^e pers. sing.	84,851	0,225	85,055
Conjugaison complète	87,515	2,220	89,503
Analyse complète	65,175	24,548	86,387

TAB. 4.13 – Performances de PCP pour les trois tâches de conjugaison. Les alphabets codant formes radicales et fléchies sont considérés comme identiques.

notre système est donc confronté au problème de « deviner » la forme abstraite (passé, radical, etc) correspondant à une chaîne phonémique donnée.

Résultats globaux Les résultats obtenus, résumés dans la table 4.13, témoignent du bien-fondé de notre entreprise, puisque non seulement le taux de silence est quasiment annulé, mais les scores en termes de correction et de justesse sont légèrement meilleurs que dans nos expériences précédentes, et très voisins des meilleurs résultats publiés dans (Ling & Marinov 1993).

On notera toutefois que les résultats plus que satisfaisants obtenus lors de l'expérience de conjugaison complète sont biaisés par la forte représentation dans nos lexiques de tests de formes radicales, qui sont de loin les plus nombreuses (ces formes représentent environ 30 % du lexique total). Pour les catégories dont la conjugaison est « difficile » (passé simple, passé composé), les résultats sont légèrement moins bons lorsque tout le système de conjugaison est appris simultanément (83,59 % de justesse pour le passé; 82,61 % de justesse pour le passé composé), par rapport à ce que l'on observe quand les conjugaisons sont apprises séparément.

Les résultats de PCP sur la tâche d'analyse sont plus décevants, notamment à cause du fort taux de silence. La tâche était sensiblement plus difficile puisque la relation entre les différentes formes d'un même verbe n'est plus directement utilisable pour la recherche de l'analogue. Le taux de justesse est néanmoins encourageant, et reflète l'adéquation des stratégies analogiques au problème de l'analyse de formes inconnues. Ce résultat demande toutefois d'être confirmé dans des conditions plus réalistes.

Examen des erreurs dans la construction du passé L'examen détaillé des erreurs produites dans la construction du passé reproduit les mêmes distributions d'erreurs que précédemment : seulement 22,78 % des verbes irréguliers sont correctement conjugués, contre plus de 93 % des verbes réguliers.

Les erreurs commises sur les verbes irréguliers sont, comme précédemment, principalement (242 sur 361) des erreurs de régularisation. On note à nouveau que dans tous les cas, les verbes sont « correctement régularisés ». L'examen des 219 erreurs commises sur les verbes réguliers se révèle plus instructive, en ce qu'elle fait apparaître deux éléments nouveaux : d'une part une très haute proportion des verbes dont le radical se termine par une occlusive dentale sont incorrectement conjugués comme les verbes pour lesquels les formes radicale et

passée sont identiques (de type *let*); d'autre part un nombre non-négligeable (42) de formes régulières sont incorrectes parce qu'elles ne respectent pas les contraintes phonotactiques de surface. On observe ainsi des formes conjuguées se terminant en /td/, /dd/, par des groupes (obstruante non-voisée + /d/); et à l'inverse des épenthèses parfaitement inutiles, comme dans *probe* conjugué /probid/, ou encore *voice* conjugué /voisid/.

La raison de ces erreurs d'un genre nouveau n'est pas difficile à comprendre, si l'on se rappelle comment fonctionne l'extraction des transformations (voir principalement la section 4.4.2). Une des conséquences immédiates de la nouvelle représentation des exemples est d'attester massivement la transformation, au niveau des formes abstraites, $+R \rightarrow +D$. À cette transformation abstraite sont indistinctement associées toutes transformations phonétiques attestées correspondant à cette flexion, aux premiers rangs desquelles naturellement, les transformations finales correspondant à la flexion régulière: $\lambda \rightarrow /d/$, $\lambda \rightarrow /t/$ et $\lambda \rightarrow /id/$; mais également d'autres, telles que $\lambda \rightarrow \lambda$ (correspondant aux verbes de type *let*, /lɛ/ → /æŋ/ (pour les verbes de type *ring*)... Confronté à une forme à fléchir, de type *radical*+*D*, PCP parvient toujours à trouver un analogue de la forme *radical*+*R*; si c'est celui-là le meilleur, il a recours aux transformations que nous venons partiellement d'énumérer pour reconstruire la forme fléchie. Quand le verbe est irrégulier, il se trouve toujours une transformation dont la partie gauche n'est pas vide, et qui s'applique en priorité: c'est un effet immédiat de l'homogénéité des micro-paradigmes. Pour les verbes réguliers par contre, les transformations avec une partie gauche vide vont être utilisées, sans qu'aucune information contextuelle ne permette de choisir laquelle est la transformation correcte; d'où les erreurs que nous venons de constater.

Cette expérience permet de mettre à nouveau en évidence une des faiblesses de PCP, à savoir l'extraction de transformations conjointes sous la forme de règles de réécriture indépendantes du contexte, qui dans ce cas précis conduisent à des erreurs grossières. Il est d'ailleurs remarquable que, compte-tenu de cette limitation, le nombre de conjugaisons réellement incorrectes soit aussi faible: quelques dizaines sur les 4000 formes que nous avons donné à conjuguer.

On notera d'ailleurs que cet effet-là est sensiblement atténué dans l'expérience de conjugaison complète. En effet, à côté de la transformation $+R \rightarrow +D$, la transformation $+Z \rightarrow +D$ est également extraite, par identification des formes correspondant respectivement à la troisième personne du singulier et au passé d'un même radical. À cette transformation « abstraite » sont associées entre autres les transformations /s/ → /t/ et /z/ → /d/ correspondant respectivement aux lois de passage pour les verbes réguliers se finissant par une occlusive¹⁹ non-voisée et voisée. En utilisant alors la forme 3^e personne comme analogue plutôt que la forme radicale, PCP est à même de retrouver correctement l'alternance /t/-/d/, qui correspond pour le passé à l'alternance /s/-/z/ pour la 3^e personne. Le problème reste bien sûr entier pour ce qui concerne la conjugaison des verbes pour lesquelles une voyelle épenthétique est requise, puisque le système de représentation utilisé, qui ne « connaît » pas le système des traits distinctifs, interdit que soit reconnue l'analogie structurelle entre les sibilantes

19. Plus précisément, par une occlusive non-dentale et non-sibilante, compte-tenu des remarques précédentes, puisque dans ces cas précis, on observe une épenthèse, soit dans la forme passée, soit dans la forme correspondant à la troisième personne.

(pour la 3^e personne) et les occlusives dentales (pour le passé).

En revanche, les résultats de cette expérience ont mis en évidence une faille dans le mécanisme analogique mis en œuvre dans PCP. Le taux de verbes irréguliers régularisés, à peu près égal à celui de l'expérience sur le seul passé montre en effet qu'un élément important de l'organisation du système de conjugaison n'a pas été saisi, à savoir la condition structurelle de paradigme, qui s'exprime par le fait qu'un verbe possédant un participe irrégulier possède également un passé irrégulier, et réciproquement. Ici encore, il semble que la sur-représentation numérique des formes radicales ait fortement contribué à masquer cette régularité, en privilégiant fréquemment la transformation $+R \rightarrow +D$ par rapport à la transformation $+P \rightarrow +D$.

4.5.3 Discussion

En conclusion, l'ensemble de ces expériences nous a permis de démontrer toute la validité de l'approche analogique pour traiter des tâches strictement morphologiques. En particulier nous avons obtenu des résultats équivalents ou supérieurs à ceux obtenus à l'aide de mécanismes d'apprentissage différents, tels que des perceptrons multi-couches, ou des arbres de décision.

Il nous semble aussi être parvenu à un système qui répond à une partie des critiques adressées par Pinker et Prince au modèle connexionniste de Rumelhart et McClelland, en ce que notre mécanisme fonctionne aussi bien dans le sens passé→présent que dans l'autre (c.f. les expériences sur la conjugaison complète), qu'il apprend l'ensemble du système de conjugaison des verbes, en tirant partie des phénomènes qui apparaissent dans des conjugaisons différentes, et qu'il ne génère pratiquement pas de forme invraisemblable. De plus, dans la mesure où de plus notre algorithme fonctionne à un niveau de représentation strictement symbolique, il est à même de tirer profit de toutes les informations symboliques supplémentaires qui lui permettraient d'améliorer les résultats; par exemple d'exploiter un mécanisme d'identification lexicale plus fin. De même, les formes produites sont directement réutilisables dans d'autres modules du traitement du langage naturel, sans requérir le moindre décodage.

Même s'il nous reste un long chemin avant d'être en mesure de proposer un modèle cognitif complet, qui d'une part mettrait cette compétence flexionnelle en relation avec les autres modules de traitement du langage naturel (module de sélection d'une forme, ...); et d'autre part permettrait de reproduire les courbes d'apprentissage et les erreurs observées chez l'enfant; il nous semble avoir démontré que l'utilisation de modèles analogiques pour la morphologie permet de construire des mécanismes capables d'exhiber en surface un comportement similaire à celui d'un système de règles.

Ces expériences nous ont également permis de montrer que l'utilisation d'analogies telles que celle qu'extrait PCP a de bonnes propriétés :

- elle offre un mécanisme capable de restituer en surface des contraintes phonologiques qui s'expriment dans les termes d'un système de traits distinctifs sans que l'on dispose de représentation de ce système de traits. C'est le sens de l'observation que les verbes régularisés le sont de manière

phonotactiquement correcte;

- elle permet de façon naturelle d'intégrer une des caractéristiques importantes des verbes irréguliers, à savoir leur mono-syllabité; là encore sans qu'à aucun moment l'importance de ce facteur-là ne soit explicité;
- elle parvient à repérer des micro-paradigmes irréguliers d'une manière qui est cohérente par rapport au système des verbes anglais;
- elle autorise à tirer un plein profit de l'organisation du système paradigmatique de conjugaison, comme en témoigne la « découverte » de l'alternance entre les suffixes à accoler aux radicaux terminant par des occlusives voisées ou non-voisées;
- elle fournit également un outil pertinent pour résoudre non seulement des tâches de génération, mais également d'analyse.

4.6 Conclusion

Nous avons présenté dans ce chapitre deux algorithmes fondés sur un même principe, celui de l'extraction et l'exploitation de liens paradigmatiques. Nous avons mis en évidence la pertinence linguistique d'une telle démarche, et montré, à partir d'applications à la tâche de transcription graphème-phonème, et à la tâche de l'apprentissage du système de conjugaison de l'anglais, que cette technique originale permettait d'obtenir, sur ces deux tâches, des performances particulièrement intéressantes, tout en exhibant des propriétés hautement désirables.

Cette approche nous paraît ouvrir des horizons de recherche particulièrement intéressants, et ce à plusieurs titres :

- **Homogénéité et transduction**: le changement de perspective ouvert par une approche à base de couplage de transformations *homogènes* apporte une solution élégante au problème de l'alignement. De nombreux problèmes qui sont posés dans le cadre du langage naturel se laissent formaliser comme un problème de transduction d'un langage dans un autre : la transcription graphème-phonème (transduction d'un langage graphémique dans un langage phonémique), l'assignation des parties du discours (transduction d'un langage de mots dans un langage de parties du discours), la traduction, etc. Une des difficultés majeures que l'on rencontre quand on aborde ces problèmes provient de la nécessité d'apprendre simultanément un principe de segmentation et un principe de classification, à travers le calcul d'une correspondance entre des unités qui sont nécessairement plus petites que les unités sur lesquels s'effectue l'apprentissage, pour conserver à la méthode des pouvoirs généralisateurs.

L'approche que nous défendons ici ignore ce problème parce qu'elle ne considère jamais de correspondance autre que *globale* entre les éléments des deux langages. Étant donné un mot et sa transcription, le seul principe d'association que nous reconnaissons est celui qui relie ces deux termes.

Cette association est irréductible à et non-décomposable en une série d'associations entre unités plus petites. Ceci a quatre conséquences importantes :

- il est particulièrement naturel, dans une telle optique, d'utiliser les informations supplémentaires qui sont éventuellement associées aux lexies des fichiers d'exemples. C'est ce que nous avons fait en utilisant les catégories morpho-syntaxiques disponibles dans le lexique BDLEX (voir la section 4.3.3), mais on pourrait aussi bien envisager de « typer » les transformations extraites d'un lexique de noms propres en utilisant des informations concernant la langue d'origine. Bien des extensions dans ce domaine sont envisageables, comme l'utilisation des fréquences en langue pour moduler l'attraction analogique exercée par tel ou tel analogue;
- la modélisation qui découle du respect du principe d'homogénéité des transductions est conforme au fonctionnement réel de la langue, dans la mesure où elle intègre l'autonomie relative des différents systèmes de signes (graphémique et phonétique), dont nous avons mentionné diverses manifestations à la section 1.1.2. Ainsi, telle modification morphologique se traduira d'une certaine manière dans un domaine, et différemment dans l'autre, parce que les contraintes qui pèsent sur l'organisations des deux systèmes sont différentes;
- les seules transductions que nous considérons sont homogènes au sein d'un langage, là où la décomposition en unités plus petites est explicitement fondée par l'existence d'un alphabet permettant de décrire les objets manipulés.

Pour prendre un exemple, considérons le mot *chasse* : dans une optique où l'on souhaiterait établir une correspondance avec sa transcription phonétique, il est indispensable de ne pas séparer le groupe *ch* qui contribue de manière indissociable à la prononciation /ʃ/. Si l'on ne considère plus que des transformations internes à l'ensemble des chaînes de caractères alphabétiques, alors *chasse* n'est que la succession des lettres qui le composent, et n'est sous aucun rapport plus proche par exemple de *chaste* que de *classe*, qui sont simplement deux voisins lexicaux à distance lexicale 1.

C'est essentiellement grâce à cette propriété importante d'homogénéité des transformations que nous avons pu appliquer, avec un certain succès, les mêmes procédures au problème de l'apprentissage de la transcription comme au problème de l'apprentissage de la conjugaison des verbes anglais. Nous envisageons d'étudier les potentialités de cette stratégie à d'autres problèmes, notamment celui de l'assignation de traits sémantiques ou de la sous-catégorisation de mots inconnus.

- cette approche semble enfin permettre de s'affranchir partiellement de la condition implicite de conservation stricte de l'ordonnancement des chaînes mises en correspondance lors de la transduction, qui sous-tend les approches classificatoires présentées au chapitre 2, comme les approches à base de recombinaison de segments développées au chapitre 3. Ce dernier point reste à vérifier sur des tâches concrètes,

mais il est remarquable que cette hypothèse ne soit pas utilisée dans notre modèle, dans lequel la seule pré-supposition qui est faite est qu'une modification initiale dans un domaine induit une modification initiale dans l'autre domaine;

Finalement, il nous semble important de souligner la convergence entre notre démarche, qui implante une forme de couplage entre deux systèmes de calcul associatif, et des développements récents dans le domaine de la traduction automatique. En particulier les approches mises en œuvre dans (Vidal, Pieraccini, & Levin 1993), qui utilisent des techniques de couplage stochastique entre deux automates à états finis, ou dans (Wu 1995), dans lequel la traduction s'effectue par couplage de formes particulières de grammaires algébriques, nous semble partager avec la nôtre un certain nombre de similarités, au premier rang desquelles la conception commune du processus de transduction comme résultant d'un couplage, et non pas d'un appariement.

- **Apprentissage et Paradigmes**: il est un second aspect de l'approche paradigmatique qui nous tient à cœur, mais que nous n'avons pas eu le temps de développer ni d'approfondir; et qui trouve son origine dans l'observation des « trous » dans les dérivations morphologiques. Nous l'avons noté, un certain nombre de raisons que l'on peut avancer pour justifier l'existence de ces trous (étymologie, incompatibilité sémantique) sont sans relation directe avec le problème de la transcription qui était le nôtre.

Cette remarque peut, à notre avis, être généralisée: les données linguistiques dont on dispose pour réaliser une tâche d'apprentissage donnée présentent des régularités qui sont orthogonales ou, du moins, sans rapport avec la tâche d'apprentissage en question. Pour donner un autre exemple de cet état de fait, il suffit de noter qu'aucun corpus naturel de l'anglais ne contiendra la phrase: *Colorless green ideas sleep furiously*, qui pourtant constitue un exemple parfaitement valide de phrase anglaise pour qui ne se préoccupe que d'en apprendre la grammaire.

Disposer d'un principe tel que le principe d'analogie paradigmatique peut être alors vu comme un moyen très général pour soit engendrer à partir d'un lexique donné un grand nombre d'éléments supplémentaires (ainsi les *gazonner*, *poisonner*, ... qui font défaut au lexique français), exemples supplémentaires qui pourront faciliter ou rendre plus robuste un apprentissage statistique; soit pour au contraire diminuer la taille de l'espace des exemples et des paramètres à estimer, en utilisant les transductions paradigmatiques comme des opérateurs de projection particuliers dans l'un ou l'autre espace.

De manière plus générale, il nous semble que l'approche paradigmatique introduit une avancée dans le domaine de l'apprentissage automatique du langage naturel, dans la mesure où elle se fonde sur une conception des exemples qui diffère profondément de celle mise en œuvre dans les démarches classificatoires. Pour résumer notre position de façon lapidaire, les lexiques de langage naturel ne sont pas des ensembles de réalisations indépendantes uniformément distribuées, mais sont structurées de manière profonde par des principes d'organisation très généraux. C'est à notre avis

en réfléchissant sur des méthodes d'apprentissage qui s'appuient explicitement sur ces structures, plutôt qu'elles les ignorent, que l'on pourra construire des systèmes réellement performants.

- **Un modèle alternatif de la lecture ?** : il nous semble enfin important de souligner, même si la plausibilité cognitive de notre modèle n'est pas ici notre préoccupation principale, que l'approche exemplifiée à travers PPP et PCP fournit un modèle formel de la prononciation par analogie, qui, à la différence des approches connexionnistes, préserve le concept de lexique, s'affranchissant ainsi d'une critique importante. À la différence des autres modélisations, elle permet de donner un sens précis au concept d'analogie, et donc de tester si les analogues qui sont extraits ont une pertinence par rapport aux observations empiriques. Enfin, elle permet de modéliser les effets bien attestés, comme l'effet de « gang », nous l'avons vu à la fois dans les expériences sur la transcription et dans celles sur les verbes anglais.

La raison de ce succès est claire : PCP tire en fait partie d'une redéfinition complète de la topologie sur l'espace des chaînes de graphèmes, en troquant la distance habituelle (insertion/suppression) par une distance définie par les transformations paradigmatiques. Cette approche permet de modéliser le fait que *manquement*, malgré une distance lexicale importante, est proche de *manquer*, puisqu'il existe une transformation massivement attestée qui permet de passer d'un terme à l'autre (et qui correspond à un schéma régulier de nominalisation). De cette manière, les voisinages dans lequel sont recherchés les analogues ont une véritable pertinence linguistique, ce qui permet de mieux contrôler les effets de « gang ».

Dans ce contexte, un examen plus poussé des capacités d'une approche paradigmatique à rendre compte d'autres faits expérimentaux (influence des fréquences, reproduction de courbes d'apprentissages naturelles, des erreurs naturelles) nous semble pouvoir se révéler également une direction d'exploration fructueuse.

Conclusion

Nous avons présenté, dans ce document, un certain nombre de techniques permettant de réaliser une tâche précise, consistant à extraire automatiquement d'un lexique de transcriptions phonétiques les connaissances permettant de nourrir un mécanisme de phonétisation donné. Notre apport principal, dans ce contexte, a été de proposer et d'évaluer de manière comparative plusieurs modélisations et mises en œuvre différentes de la manière dont il est possible de tirer parti de ressemblances ou d'analogies entre des termes dont la prononciation est inconnue et des lexies.

Nous avons en particulier montré, en nous appuyant sur l'exemple précis d'un algorithme de transcription par induction d'arbre de décision, DEC (Torkolla 1993), qu'un certain nombre de critiques étaient opposables aux approches fondées sur un principe classificatoire. En particulier, nous avons mis en évidence leur insuffisante capacité à tirer parti efficacement de gros lexiques d'exemples, les problèmes posés par la génération de transcriptions multiples, et souligné l'importance des données de supervision dans ce contexte, ainsi que les problèmes que soulevait leur production.

Nous avons ensuite montré comment il était possible de concevoir le mécanisme de transduction sous la forme d'un principe de concaténation de larges segments. En étendant l'algorithme de Dedina et Nusbaum (1991), dont nous avons mis en évidence l'insuffisante capacité de généralisation, nous sommes parvenu à des performances en termes de transcriptions correctes supérieures à celles obtenues en utilisant une approche classificatoire. Nous avons finalement mis en évidence les convergences profondes entre cette ligne de recherche et celle développée indépendamment dans le cadre de la modélisation multigramme, et suggéré des voies d'amélioration de ces techniques, dont le champ d'application excède de loin la tâche de transcription graphème-phonème. Ces divers développements nous ont toutefois convaincu que la modélisation des processus de lecture par analogie par des mécanismes fondés sur la concaténation de segments était inappropriée, dans la mesure où il était impossible de définir dans ce cadre les notions essentielles de *voisinage lexical* et de *mot analogue*, et d'explicitier de manière transparente les processus de transfert analogique.

Nous avons finalement conçu un algorithme, linguistiquement fondé sur le concept de paradigme, permettant d'apporter un certain nombre de réponses aux critiques développées à l'encontre des techniques classificatoires. Ce principe de transcription entièrement original, qui procède par extraction et utilisation d'un couplage entre systèmes de calcul associatif dans deux domaines distincts,

nous a permis de modéliser de manière convaincante le mode d'interaction complexe entre les systèmes graphémiques et phonétiques, et notamment l'absence d'assujettissement d'un système de représentation à l'autre.

Cette modélisation a conduit la reformulation de l'analogie comme un principe de similarité dans une topologie émergeant simultanément dans les domaines graphiques et phonétiques de l'interaction entre les éléments dans ces deux domaines. En effet, le mécanisme d'extraction de transformations paradigmatiques peut être reformulé comme l'extraction d'une distance particulière particulière sur les ensembles de chaînes. Cette distance permet de construire une nouvelle topologie qui, à l'inverse de celle induite par la distance d'édition, est respectueuse de la forme particulière que prennent les relations lexicales dans une langue donnée. Cette topologie, qui repose simultanément sur un principe formel (la condition d'alternance) et sur une mesure fréquentielle, semble offrir un fondement plus solide pour modéliser les phénomènes d'influence lexicale observés lors d'expériences de lecture à voix haute. En particulier, elle offre une définition fondée au concept de « voisinage lexical », ce qui permet d'envisager de tester précisément, en comparant ces voisinages avec ceux que définiraient des humains, ses capacités de prédiction.

Du point de vue des résultats, nous avons pu obtenir, grâce à cette technique, des performances de transcription particulièrement satisfaisantes, compte-tenu de l'abandon des données de supervision (alignement). Deux autres caractéristiques intéressantes de cet algorithme ont également pu être mises en évidence, à savoir sa capacité à identifier de manière fiable les chaînes graphiques atypiques par rapport à un lexique donné, et à tirer parti de manière efficace des vastes bases de données dont nous disposons. Il nous semble donc qu'un tel modèle répond de manière convaincante aux exigences que nous nous étions fixées en débutant ce travail, et fournit en particulier, pour diverses raisons, des éléments de solution particulièrement adaptées au problème de la transcription des noms propres.

L'applicabilité d'une telle approche à d'autres problèmes délicats du traitement du langage naturel a été démontrée dans le cadre d'une expérience d'apprentissage du système de conjugaison de l'anglais. Cette expérience a également permis de mettre en évidence de manière plus nette la capacité de notre modèle à exhiber en surface un comportement très voisin de ce que produirait un système de règles, tout en conservant une flexibilité importante permettant de rendre compte de l'inégale généralité et applicabilité des « règles » morphologiques (par exemple le fait que les verbes réguliers improprement irrégularisés sont préférentiellement mono-syllabiques). Ces expériences ont également démontré que notre modèle était capable d'obtenir, pour la tâche de conjugaison, des performances en généralisation voisines ou supérieures à celles de systèmes classificatoires.

Enfin, nous avons suggéré diverses extensions de notre modèle, qui permettraient d'améliorer les performances en termes de temps de traitement et de correction des transcriptions. Sans préjuger de la complexité de mise en œuvre ces divers amendements, ni de leur efficacité, il nous semble que la façon dont nous avons pu les exprimer reflète une caractéristique importante de notre modèle, à savoir sa capacité à gérer simultanément des contraintes symboliques (par exemple les catégories morpho-syntaxiques des lexies, ou encore une gram-

maire phonotactique), et numériques (par exemple les fréquences des lexies, ou encore des modèles de langage).

D'une manière plus générale, une contribution importante de notre travail nous paraît être une certaine clarification du concept d'analogie, et de la manière dont, dans le cadre précis qui nous concerne, les analogies peuvent « fonctionner ». Nous partions en effet d'une situation dans laquelle ce terme était utilisé pour qualifier :

- un principe de fonctionnement cognitif observé dans diverses situations expérimentales de lecture (Glushko);
- des systèmes de transcription fondés sur le recollement de segments de lexies (Byrd et Chodorow, Dedina et Nusbaum);
- des systèmes de transcription fondés sur une classification utilisant des fonctions de similarité/dissimilarités locales (Skousen, Stanfill et Waltz) ou globales (réseaux connexionnistes);
- un facteur d'évolution diachronique et de productivité synchronique dans la tradition linguistique (Saussure);
- un transfert structurel de propriétés et de relations entre éléments de deux systèmes d'objets (au sens large).

Nos travaux nous ont finalement conduit d'une part à disqualifier et les systèmes classificatoires et les approches par recollement de segments (pour ces dernières, non pas au regard de leurs performances, mais bien au titre de la modélisation qu'elles proposent de la transcription par analogie), et d'autre part à réconcilier les interprétations linguistiques et logiques dans un modèle unifié permettant de rendre compte d'un certain nombre d'observations faites par les psychologues.

À l'issue de ce nécessaire travail de clarification, nous nous trouvons dans une posture plus confortable pour envisager les possibles perspectives ouvertes par nos recherches. Nous avons déjà eu l'occasion de souligner, à divers endroits, les directions d'approfondissement qu'il nous semblait fructueux d'explorer. De manière plus générale, il nous semble que l'extension des différentes techniques que nous avons pu proposer doit passer par une reformulation plus formelle dans le cadre de la théorie des langages.

Nous avons suggéré comment cette extension peut être envisagée en ce qui concerne les approches par recollement de segment: les considérer comme des procédés d'instanciations particuliers de grammaires régulières sur les segments offre de nombreuses perspectives, dans la mesure où les automates à états finis, sous laquelle elles sont manipulées en machine ont été l'objet de nombreuses études. Il est donc envisageable de mettre en œuvre toute une batterie de procédures, qui permettraient de les déterminer et de les minimiser (Hopcroft & Ullman 1979), d'en inférer des grammaires plus générales (Miclet 1979), ou encore de procéder à des extensions stochastiques.

En ce qui concerne les approches fonctionnant par extraction de couplages de transformations, le préalable à des extensions significatives semble être une

reformulation dans un cadre formel plus contraint que celui fourni par le calcul associatif, dans lequel se posent, nous l'avons vu, des problèmes algorithmiques complexes, que l'utilisation d'une mesure de probabilité sur les transformations ne parvient qu'imparfaitement à faire disparaître. Une fois ce travail de reformulation effectué, il deviendra d'une part possible d'envisager l'application de ce modèle à d'autres tâches d'extraction de connaissances lexicales pour le traitement du langage naturel, et de conduire une procédure de validation plus complète du modèle cognitif de lecture par analogie, et plus généralement du traitement lexical, qu'il met en œuvre.

Annexe A

Les mécanismes d'alignement

Après avoir présenté le problème de l'alignement automatique de couples (graphie, transcription), nous décrirons quelques unes des différentes approches proposées pour résoudre ce problème, avant de détailler notre propre algorithme d'alignement, ainsi que ces performances sur le lexique NETTALK, qui est le seul pour lequel nous disposons d'un alignement bi-univoque manuel.

A.1 Position du problème

On désigne sous le terme général d'*alignement* la mise en correspondance de deux séquences d'événements. Cette opération a de multiples applications en traitement automatique du langage naturel et de la parole : alignement (des phrases) de corpus bilingues pour la traduction automatique, alignement d'un signal temporel de parole avec la suite de phonèmes correspondante..., sont ainsi nécessaires pour constituer des corpus d'apprentissages dans différents domaines.

Dans le cadre qui nous préoccupe, l'apprentissage de systèmes de transcription graphème-phonème, aligner signifie mettre en relation des séquences de graphèmes avec les séries correspondantes de phonèmes. La mise en correspondance peut être bi-univoque (à chaque symbole d'une chaîne correspond un symbole d'une autre), ou plus lâche, en acceptant la mise en correspondance de séquences de symboles (nous référerons à ces alignements comme étant *plurivoques*).

La figure A.1 présente ainsi différents alignements possibles du mot *voyager* avec la transcription phonémique correspondante /vwajaze/.

Dans le cas d'alignement bi-univoque, tels que dans les exemples (A.1.a) et (A.1.b), la différence de longueur entre les deux chaînes conduit à mettre en correspondance certains des symboles avec un symbole vide fictif, figuré ici par un /-/. Dans le cas A.1.a, on résout le problème de l'alignement de la séquence *oy* et de la séquence /way/ en introduisant un nouveau symbole phonétique,

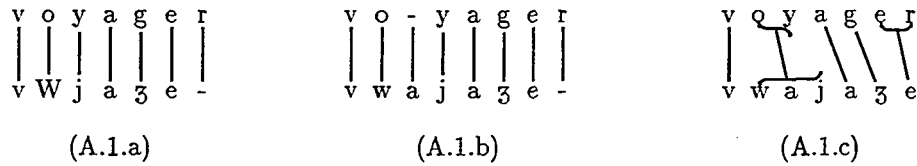


Figure A.1: Différents alignements de *voyager* avec */vwajaze/*.

/W/, qui représente le diphone */wa/*. Dans le cas A.1.b, ce problème est résolu en introduisant un caractère graphémique nul, représenté par *-*.

La procédure d'alignement implique donc, en général, un bon nombre de décisions arbitraires de la part de « l'aligneur ». En effet, des langues telles que le français, l'anglais ou l'allemand admettent des digraphes (séquence de deux graphèmes) qui correspondent (presque) systématiquement à un phonème unique (ainsi *ch*, qui, prononcé en français */ʃ/*, */k/*, et marginalement */x/* dans certaines régions, ne correspond jamais à plus d'un phonème) ou au contraire, des graphèmes susceptibles de correspondre à plusieurs phonèmes (ainsi la lettre *x*, susceptible de se prononcer en français */ks/* ou *gz*).

L'alignement plurivoque, s'il permet d'éviter ce type de problème, semble d'une part moins bien adapté aux systèmes d'apprentissage fondés sur des modèles de classification (réseaux de neurones, construction d'arbres de décision, etc), et nécessite pratiquement la fixation arbitraire des tailles maximales de séquences admises. En l'absence d'un tel seuil, en effet, l'alignement plurivoque optimal, est en effet celui qui met simplement en correspondance chaque graphie avec la transcription correspondante.

A.2 Quelques méthodes d'alignement

A.2.1 Alignement à base de règles

Cette méthode d'alignement plurivoque a été proposée par Lawrence et Kaye (Lawrence & Kaye 1986), dans le contexte de la mise en forme d'un dictionnaire phonétique.

Algorithme

La première étape du traitement consiste à dresser, pour chaque séquence de phonèmes $\phi_{1,k}$ d'un alphabet prédéterminé *A* la liste des graphèmes $C(\phi_{1,k})$ susceptibles de lui correspondre. Lawrence et Kaye incluent par exemple les séquences suivantes dans leur table (voir la table A.1).

Une fois ce recensement effectué, pour chaque couple (graphie $g_{1,n}$, phonétique $\pi_{1,p}$) en entrée, l'algorithme procède de la façon suivante. En traversant la chaîne phonétique de la gauche vers la droite, on cherche la plus longue sous-chaîne $\pi_{i,i+k}$ de $\pi_{1,p}$ présente dans *A*. On cherche alors si il existe *l* tel que $g_{j,l}$

$\phi_{1,k}$	$C(\phi_{1,k})$	Comme dans ...
/aɪ/	y	by
/aɪ/	uy	buy
/aɪ/	igh	might
/aɪ/	ey	eyes

TAB. A.1 – *Quelques correspondances graphémiques de la séquence /aɪ/*

est dans $C(\pi_{i,i+k})$. Si c'est le cas, on continue en examinant $\pi_{i+k,p}$ et $g_{j+l,n}$. Sinon, on essaie de remettre en cause k en choisissant une chaîne phonétique plus courte. Si c'est impossible, alors on essaie de remettre en cause des $\pi_{i',i'+k'}$ précédemment sélectionnés. Si enfin cette dernière tentative échoue, alors le couple d'entrées n'est pas alignable à l'aide de la table de correspondance pré-définie.

Discussion

Cette méthode présente plusieurs avantages. Tout d'abord elle semble fournir de très bons résultats aux procédures de tests décrites par ces auteurs. Ensuite, elle permet de contrôler directement les séquences alignables entre elles (ce sont très précisément celles disponibles dans la table de correspondance, à l'exclusion de toute autre).

Le prix à payer est bien entendu celui de la construction de la table (celle de Lawrence et Kaye contient quelque centaines de correspondances), ce qui nécessite l'intervention d'un expert, et probablement, si les lexiques sont assez volumineux, un certain nombre d'essais et d'erreurs avant d'arriver à une table capable d'aligner tous les mots (on n'aura jamais de garantie de pouvoir aligner un mot inconnu).

Luk et Damper (Luk & Damper 1992; 1993) ont mis au point des techniques pour extraire automatiquement de lexiques des jeux de correspondance tels que ceux nécessaires à l'algorithme de Lawrence et Kaye, et ont montré que les alignements dérivés des tables de correspondance automatiquement construites étaient comparables à ceux obtenus à l'aide de tables construites manuellement. La construction automatique des tables de correspondance remet toutefois en cause le contrôle sur les séquences alignables entre elles, ce qui n'était pas le moindre des intérêts de cette méthode à base de règles.

A.2.2 Alignement par segmentation d'image

Description de la Méthode

Pour s'affranchir de la nécessité d'une intervention manuelle pour aligner automatiquement graphèmes et phonèmes, Luk et Damper (Luk & Damper 1991) ont proposé une méthode d'alignement multivoque originale, inspirée des méthodes utilisées pour segmenter des images.

Le principe de leur méthode est le suivant. La première étape consiste à

acquérir à partir d'un lexique de couples (graphie, phonétique) non alignés des indices d'association entre les différents symboles. Les indices calculés visent à rendre compte du degré d'association entre phonèmes et graphèmes, et également entre graphèmes entre eux, et phonèmes entre eux. L'indice d'association de deux symboles est une fonction de leur position relative dans les chaînes du lexique d'exemples. Soient ainsi deux symboles s_1 et s_2 dans $G \cup P$, et n un entier, l'indice mesurant l'association de s_1 et s_2 quand ils sont séparés de n caractères est simplement :

$$a(s_1, s_2, n) = p(s_2, n | s_1) \quad (\text{A.1})$$

Cet indice est estimé par les fréquences relatives (éventuellement pondérées par les fréquence en langue) de ces événements sur un gros lexique.

Une fois ces indices acquis, Luk et Damper présentent deux stratégies pour dériver les alignements. La première de ces méthodes consiste à considérer chaque couple à aligner $(g_{1,n}, \pi_{1,p})$ comme une matrice $n \times p$ d'indices d'association dont le terme général $I_{i,j}$ vaut $a(g_i, \pi_j, i - j)$. L'alignement se déduit de I par transformation de I en une matrice booléenne B (la matrice d'adjacente du graphe représentant les alignements). La transformation appliquée à I consiste remplacer par 1 tous les $I_{i,j}$ qui sont supérieurs à un seuil S , et par 0 tous les autres. Le seuil S est choisi comme étant le plus petit des maximums en ligne et en colonne de I , de manière à garantir que chaque lettre soit associée à un phonème, et réciproquement.

La seconde consiste à construire le graphe symétrique complet dont les sommets sont les graphèmes et les phonèmes, les arcs étant valués par les indices d'association. L'alignement résulte (dans les cas favorables) de l'extraction de ce graphe d'une partition des sommets en arbres de poids minimum. Cette dernière procédure, si elle présente l'avantage de tirer parti des indices d'association entre graphèmes et entre phonèmes, ne garantit pas que chaque graphème (respectivement phonème) sera mis en relation avec un phonème (respectivement graphème). Il est donc malheureusement possible d'obtenir par cette méthode des alignements incomplets.

Discussion

Luk et Damper ne donnent pas de résultats chiffrés, ce qui rend cette méthode difficilement évaluable. L'idée de tenir compte simultanément de tous les critères d'associations disponibles est néanmoins séduisante, puisqu'elle permet de mettre en évidence les influences entre graphèmes qui conduisent à un alignement donné. Ainsi, dans le cas du couple (*make*, /meik/), la partition du graphe complet en arbre de poids minimum contient une arête entre les sommets "a" et "e", en plus de l'arête d'alignement entre a et /ei/, témoignant de l'influence du e sur cette prononciation particulière du a. L'abandon de cette technique d'alignement dans les travaux ultérieurs de ces deux auteurs ((Luk & Damper 1992; 1993)) jette néanmoins un doute sur l'efficacité réelle de cette méthode.

On peut légitimement penser que le grand nombre d'indices d'associations à estimer (ils se chiffrent en dizaine de milliers), indices qui sont de surcroît

« spatialement » neutres, dans la mesure où ils ne prennent en compte que des positions relatives dans des chaînes qui peuvent être de longueurs différentes, introduisent beaucoup de bruit, et nuisent à la robustesse de la méthode.

A.2.3 Alignement à base de modèles markoviens

De nombreux auteurs ont enfin fait appel pour aligner des chaînes de graphèmes avec leur transcription phonétique à une classe de modèle massivement utilisée pour le traitement de la parole, à savoir les modèles de Markov cachés (voir la présentation à la section 2.2.3).

À partir de ce cadre très général, plusieurs approches sont permises. On peut par exemple, à l'instar de ce que propose van Coile (1991), identifier l'ensemble des symboles en sortie à l'alphabet graphémique, et construire pour chaque phonème le modèle à cinq états suivant :

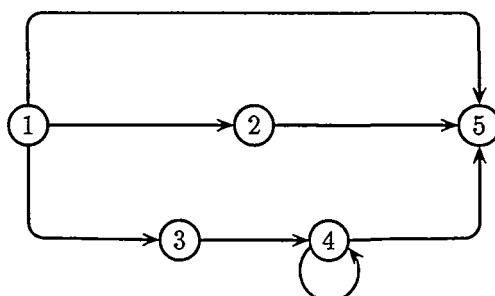


FIG. A.2 – *Modèle de phonème pour l'alignement*

Les états d'un modèle correspondant à un mot donné seront alors obtenus en identifiant l'état 1 du phonème π_i à l'état 5 du phonème π_{i-1} .

On peut également, plutôt que travailler sur l'alphabet des graphèmes, prendre comme ensemble de symboles en sortie l'ensemble des chaînes de graphèmes de taille inférieure à une taille (raisonnable) fixée à l'avance. C'est la démarche prise par van Coile dans des travaux plus récents ((van Coile 1993)), qui, comme la démarche précédente, conduit à la construction d'alignements plurivoques.

Dans les deux cas, le grand nombre de paramètres à estimer conduit à des problèmes d'estimations bien connus pour les probabilités faibles, qui correspondent à des séquences dont rien ne garantit qu'elles seront observées sur le lexique.

Discussion

Ces méthodes d'alignements présentent l'avantage d'être relativement souples au niveau du choix des alphabets et des états, d'être (relativement) faciles à mettre en œuvre à l'aide d'algorithmes connus (algorithme de Viterbi ou de Baum-Welsh), et de donner des résultats très satisfaisants.

Elles sont par contre assez sensibles à la manière dont les algorithmes sont initialisés, ce qui requiert parfois l'ajout de quelques connaissances symboliques (par exemple la définition a priori de quelques alignements atomiques privilégiés) pour pouvoir reproduire les résultats obtenus manuellement (par exemple dans (van Coile 1993)).

En effet, une fois le modèle spécifié, le modèle va fonctionner de façon complètement non supervisée, et peut donc, suivant le lexique et les valeurs initiales, parvenir à des résultats qui ne seront pas toujours satisfaisants ou cohérents pour un linguiste. Ceci tout particulièrement si l'on cherche à obtenir des alignements bi-univoques.

Prenons un exemple parlant pour le français, à savoir l'alignement du phonème / $\tilde{a}$ / et des digraphes *an*, *en*, *am* et *em*. Il semble naturel, pour des phonéticiens du moins, de toujours aligner / $\tilde{a}$ / avec la voyelle du bigramme correspondant, en premier lieu parce que la règle phonologique qui s'applique est la nasalisation d'une voyelle, et non pas la vocalisation d'une consonne nasale, ensuite parce qu'il semblerait de toute façon étrange d'aligner une consonne graphique sur une voyelle phonétique.

Pourtant, suivant les algorithmes et les proportions de ces quatre bigrammes dans le lexique d'apprentissage (ainsi que suivant les conditions initiales, et quelques autres paramètres), la meilleure solution peut-être l'alignement de la voyelle phonétique et de la consonne nasale graphémique, ce qui est pour le moins assez contre-intuitif.

A.2.4 Algorithme d'alignement

Le modèle

Le modèle (simple) sur lequel nous nous sommes fondés est également descriptible dans les termes généraux d'un modèle de Markov caché. Étant donnée une chaîne $g_{1,n}$ de graphèmes, on suppose que cette chaîne est émise par un processus markovien dont les états (cachés) sont simplement les phonèmes. Formellement, en conservant les notations précédentes, notre modèle se définit par :

- $S = \{s_1\} \cup \{\text{ensemble des phonèmes}\} \cup \{s_\emptyset, \text{correspondant à un phonème nul}\}$
- s^1 = état fictif : début de chaîne.
- $W = \{\text{ensemble des graphèmes, plus un graphème nul}\}$
- $T = \{t_{i,j,k}, \text{pour tout } (s_i, s_j) \text{ dans } S \times S, \text{ et pour tout } w_k \text{ dans } W\}$, les seules transitions $t_{\emptyset, \emptyset, k}$ étant interdites (de probabilité rigoureusement nulle pour tout k).

Étant donné ce modèle, l'alignement d'une chaîne graphique $g_{1,n}$ avec la chaîne phonétique correspondante sera donné par la séquence des paires (g_t, π_{t+1}) telle que

$$P(g_{1,n}, \pi_{1,n}) = \text{Max}_{s_{1,n}} P(g_{1,n}, s_{1,n})$$

Il est à noter qu'aussi bien dans la liste des symboles émis que dans la liste des états figure un symbole nul. Afin de contrôler l'émission de graphèmes par des phonèmes nuls, qui correspondent en fait à l'alignement de plusieurs phonèmes sur un même graphème (cas de x , etc), on interdit que deux états correspondants aux phonème nul se succèdent. C'est le sens de notre restriction sur l'ensemble des transitions. Cette possibilité d'émettre des graphèmes nuls introduit d'ailleurs une incertitude sur la chaîne émise, la séquence graphémique pouvant se révéler être une séquence quelconque bâtie en insérant des graphèmes nuls dans $g_{1,n}$.

L'algorithme

Dans la mesure où le nombre de paramètres à estimer du modèle ainsi spécifié reste grand ($|S|^2 \times |W|$), nous avons fait l'hypothèse (classique) supplémentaire que $\forall (s_i, s_j) \in S \times S, p_{i,j,k} = p_{i,k}$, en clair que la probabilité d'émission d'un symbole ne dépend que de l'état de départ (et pas de l'état d'arrivée). Cette hypothèse permet de limiter les problèmes notoires que posent dans ce type de modèles l'estimation des probabilités faibles (correspondant à des transitions/émissions jamais ou rarement observées). Les probabilités sont estimées itérativement par l'algorithme de Viterbi. La procédure complète est résumée dans l'algorithme 8.

Algorithme 8: Algorithme d'alignement

Entrée: Un lexique $\mathcal{L}$ de formes non-alignées

Sortie: Un lexique $\mathcal{L}$ de formes alignées

ALIGNER($\mathcal{L}$)

- (1) Initialisation des probabilités
- (2) **tant que** les probabilités ne sont pas stables
- (3) **pour tout** $(x, \phi(x)) \in \mathcal{L}$
- (4) calculer un alignement compte tenu des probabilités
- (5) mettre à jour les fréquences d'alignements atomiques
- (6) **fin**
- (7) mettre à jour une nouvelle les probabilités
- (8) **fin**
- (9) **pour tout** $(x, \phi(x)) \in \mathcal{L}$
- (10) calculer un alignement compte tenu des probabilités finales
- (11) en déduire itérativement la qualité de l'alignement
- (12) **fin**

La qualité intrinsèque d'un alignement d'un lexique d'exemple est mesurée dans notre modèle par la probabilité moyenne d'émission des mots du lexique compte-tenu des probabilités. Pour s'affranchir des problèmes numériques liés à la représentation en machine des très petites probabilités, nous avons en fait considéré (au signe près) la moyenne des logarithmes des probabilités. Notre critère pour évaluer un jeu de probabilité (et l'alignement afférent de N exemples) vaut donc :

$$C = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^{n_i} -\log(pr(g_{i,j}|\pi_{i,j})) \quad (\text{A.2})$$

Bruitage

Complètement spécifié, le problème à résoudre (minimiser C) est un problème d'optimisation combinatoire. Il était tentant d'essayer d'appliquer des techniques classiques en optimisation combinatoire afin de se prémunir contre la convergence de notre algorithme vers un optimum local.

L'idée générale des techniques (recuit simulé ou bruitage) est de ralentir la convergence de l'algorithme vers un optimum local, soit en refusant (aléatoirement) de choisir la meilleure solution afin de se donner le temps d'explorer d'autres régions de l'espace de recherche (c'est l'idée générale du recuit simulé); soit en perturbant aléatoirement les solutions d'une itération (c'est l'idée générale du bruitage).

Nous nous sommes orientés vers le bruitage (Charon & Hudry 1993) essentiellement pour sa facilité de mise en œuvre : utiliser le recuit simulé nous aurait contraint à conserver, pour un mot donné, tous les alignements possibles, afin de pouvoir conserver des solutions suboptimales. L'intégration du bruitage dans notre algorithme était au contraire très aisé, et il a suffi d'incorporer une fonction chargée de perturber les fréquences d'alignement avant de réestimer les probabilités.

Le bruitage se caractérise par trois paramètres :

- l'amplitude relative A de la perturbation des paramètres (des fréquences dans notre cas).
- le pas de décroissance P de la perturbation (nécessaire à la convergence).
- la fréquence d'application I des perturbations (on peut ne pas vouloir perturber à chaque itération)

Pour donner un exemple, l'utilisation du jeu de paramètres ($A = 50, P = 2, I = 2$) signifie que toutes les deux itérations, on remplace les valeurs des probabilités par les valeurs bruitées, en diminuant progressivement l'amplitude du bruitage depuis la valeur 50 % par pas de 2.

La mise en œuvre d'une technique de bruitage ne modifie que marginalement l'algorithme 8. En effet, il suffit de faire suivre l'étape de ré-estimation des probabilités à l'itération n par une étape supplémentaire lors de laquelle chaque probabilité p est remplacée par une valeur aléatoire tirée uniformément dans $[max(0, (p - ((A - nP) \times p)/100), min(1, p + ((A - nP) \times p)/100)]$.

A.3 Quelques expériences d'alignement du lexique nettalk

Un grand nombre d'expériences ont été conduites pour déterminer le jeu de paramètres le plus adéquat. Ceci nous a permis de constater que plus la perturbation initiale était forte, et plus le pas de décroissance était faible, meilleur était (statistiquement) le résultat final.

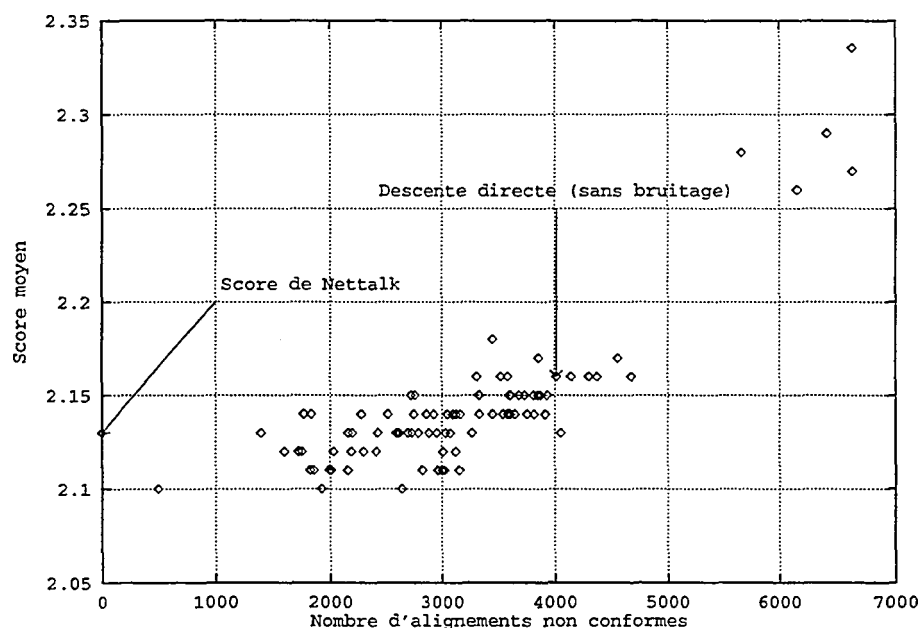


FIG. A.3 – Le critère d'alignement en fonction du nombre d'erreurs. Les données sont celles du lexique NETTALK

Nous n'entrerons pas plus avant dans le détail précis des résultats de ces expériences, dans la mesure où deux faits importants se sont révélés, qui nous ont convaincus de la vacuité de poursuivre plus avant ces expériences :

- l'alignement manuel fourni avec le lexique NETTALK, qui en dernière instance représente le résultat que nous souhaiterions obtenir automatiquement, la recherche du minimum du critère A.2 n'étant qu'un moyen de parvenir à ce résultat, est loin de correspondre à un minimum de C . Plus généralement, la corrélation entre la valeur de C et le nombre de déviations par rapport à l'alignement manuel étant relativement lâche. Ceci est particulièrement net sur le graphique reproduit sur la figure A.3, où la valeur de C est représentée comme une fonction du nombre d'erreurs d'alignement.
- DEC semble être relativement peu sensible à la façon dont les données d'apprentissage sont alignées. Diverses expériences avec des données alignées manuellement et automatiquement ont en effet fourni des résultats très similaires. Dans la mesure où le propos de notre étude concernait plus spécifiquement l'examen des mécanismes d'apprentissage, et non pas la production d'alignements optimaux ou linguistiquement pertinents, nous avons donc considéré qu'il était vain, dans cette optique, de poursuivre ces expériences.

En conséquence, pour tous les lexiques pour lesquels aucun alignement bi-

univoque manuel n'était disponible, nous avons eu recours au programme décrit dans les lignes précédentes, avec des paramètres de bruitage égaux à $A = 100$, $P = 5$, $I = 2$, paramètres qui avaient conduit à des résultats satisfaisants sur le lexique NETTALK.

Annexe B

DEC : un système classificatoire prototypique

Dans ce chapitre, nous présentons de manière complète le système de transcription qui sera utilisé dans la suite comme le système classificatoire de référence. Ce système se fonde sur l'algorithme DEC (Dynamic Expanding Context) originellement présenté par Kohonen (1986), et appliqué à la transcription graphème-phonème par Torkolla (1993). À nos yeux, ce système constitue un digne représentant de la classe des systèmes classificatoires, et présente le mérite de pouvoir être mis en œuvre de façon assez simple sans introduire de biais au niveau de la représentation des données. Dans un premier temps, nous exposerons notre implantation de cet algorithme, en insistant tout particulièrement sur les points de divergence entre notre interprétation et celle de Torkolla, avant de donner un aperçu complet des performances de ce système.

B.1 Un exemple introductif

L'idée que DEC met en œuvre est la suivante : étant donné un lexique de mots dont la transcription est connue, il s'agit de construire un arbre de classification capable de prédire *sans erreur* pour chaque lettre du lexique le phonème qui lui est effectivement associé. Cette classification s'appuie sur le contexte graphémique autour de la lettre à transcrire : on peut d'ailleurs aussi bien se représenter cet arbre de classification comme un ensemble ordonné de règles dépendant du contexte. Dans la mesure où une telle prédiction est d'une difficulté variable, et peut nécessiter suivant les cas un contexte plus ou moins long, pouvant aller jusqu'au mot entier, DEC va tirer partie d'une stratégie prudente, qui le distingue par exemple de Nettek (voir 2.2.1). Ainsi, plutôt que de se fixer une taille de contexte *a priori*, la taille du contexte est fixée librement, et n'augmente *que si cela s'avère indispensable*. Bien évidemment, cette stratégie présuppose que soit déterminée une politique pour procéder à cette expansion. À la différence d'ID3, qui calcule automatiquement le test à effectuer à chaque nœud de l'arbre, DEC requiert que soit établi par avance un schéma d'expansion du contexte (noté

Exemple	SEC	marquer ...	ajouter ...
1ère Itération			
an / ā-	-	R_0	$R_1 : a \rightarrow \bar{a} / _$
ane / an-	a	R_1	$R_2 : a \rightarrow a / _ n$
ange / ā-ʒ-	an	R_2	$R_3 : a \rightarrow \bar{a} / \# _ n$
2ème Itération			
an / ā-	#an	-	-
ane / an-	#an	R_3	$R_4 : a \rightarrow a / \# _ ne$
ange / ā-ʒ-	#an	-	$R_5 : a \rightarrow \bar{a} / \# _ ng$
3ème Itération			
an - / ā-	#an	-	$R_6 : a \rightarrow \bar{a} / \# _ n\#$
ane - / an-	#ane	-	-
ange / ā-ʒ-	#ang	-	-
Désambiguïsation fréquentielle			
-	-	-	$R_1 : a \rightarrow \bar{a} / _$
-	-	-	$R_2 : a \rightarrow \bar{a} / a _ n$
-	-	-	$R_3 : a \rightarrow \bar{a} / \# _ n$

TAB. B.1 – Apprentissage par DEC des règles de lecture de la lettre a

SEC dans ce qui suit). Nous aurons l'occasion de revenir par la suite sur le rôle joué par ce SEC, et sur les procédés qui permettent de le choisir habilement.

Le fonctionnement de DEC est décortiqué dans le petit exemple reproduit table B.1, qui décrit le processus d'extraction des règles de prononciation de la lettre a à partir d'un lexique de trois mots: (an, ane, ange). Le processus est initialisé avec une seule règle R_0 , qui dit en substance: « en l'absence de contexte, ne rien produire ». Par ailleurs, les consignes pour étendre le contexte sont de considérer d'abord la lettre à transcrire, puis alternativement une lettre supplémentaire vers l'avant puis une vers l'arrière, jusqu'à épuisement du stock de lettres disponibles.

Quand le premier mot: an est rencontré, le système construit une première règle sans contexte R_1 , qui rend compte de la prononciation du a de an en /ā/. Cette règle s'applique lorsque le mot suivant est présenté, malheureusement de façon erronée, puisque le phonème correct est un /a/. Une nouvelle règle R_2 est alors créée, par expansion du contexte selon le schéma prédéfini (i.e. en rajoutant dans le contexte le caractère suivant le a), et la règle R_1 est marquée comme ayant été prise en défaut. Ce mécanisme est réitéré tout au long de l'apprentissage, qui comprend plusieurs lectures du jeu d'exemples, jusqu'à ce que plus aucune règle ne soit prise en défaut, sauf celles pour lesquelles le contexte ne peut être étendu (ce qui se produit lorsque des homographes hétérophones se trouvent dans le lexique).

Un dernier traitement est finalement requis, qui consiste à calculer sur des bases fréquentielles la production qui doit être associée aux règles que l'on a du remettre en cause (R_0, R_1, R_2 dans notre exemple): pour celles-ci, c'est le phonème le plus fréquemment associé aux règles correspondant aux niveaux inférieurs de l'arbre qui sera produit; dans notre cas ce sera donc /ā/, qui apparaît deux fois dans notre lexique contre une seule occurrence de /a/. Cette dernière

étape garantit qu'en l'absence d'un contexte suffisant pour parvenir à une feuille de l'arbre de décision, c'est le segment le plus probable qui sera émis.

Notons que ceci permet d'associer de manière naturelle un score à chaque règle, qui témoigne de la confiance relative que l'on accorde au résultat de son application. Dans notre implémentation de l'algorithme, ce score n'est pas directement la fréquence relative de succès de la règle, mais la borne inférieure de l'intervalle de confiance associé à l'estimation de ce score interprété comme une probabilité de réussite¹. Ce mode de calcul permet de ne pas affecter le même score à deux règles qui auraient des taux de réussite égaux, mais dont l'un aurait été calculé sur un grand nombre de lexies, et l'autre sur une proportion beaucoup plus faible. Dans notre petit exemple, tous ces scores seraient donc nuls, reflétant le peu de confiance que l'on accorde à des scores calculés à partir de si peu d'exemples. On notera que l'utilisation de statistiques pour désambiguïser les contextes ambigus introduit une capacité de généralisation qui est absente d'ID3, du moins dans sa version de base. Il y a donc généralisation dans DEC d'une part parce que l'association non-ambiguë d'un contexte et d'une production est supposée vraie pour tous les mots dans laquelle cette combinaison apparaîtra, et d'autre part parce que si une telle association est ambiguë, le phonème le plus fréquent est toujours produit.

Pour en terminer avec ce point, illustrons-le par un exemple, et supposons que le mot inconnu : *ans* soit proposé au système préalablement nourri du petit lexique précédent. L'expansion maximale de contexte trouvée dans l'arbre pour ce mot est *#an*, ce qui conduit à l'application de la règle R_3 , et la production du phonème correct : /*ā*/.

DEC présente la caractéristique intéressante d'être, par construction, doté d'une mémoire infallible. En d'autres termes, une fois qu'une lexie a été apprise, sa prononciation ne sera pas oubliée, puisque tout le contexte nécessaire à la prédiction exacte de tous ses phonèmes est enregistré dans l'arbre.

Par contre DEC présente deux faiblesses manifestes par rapport aux algorithmes de la famille ID3. La première concerne la durée de l'apprentissage : alors qu'ID3 ne requiert qu'une passe de traitement des lexies, DEC demande plusieurs passages ; en fait si la plus grande taille de contexte nécessaire pour désambiguïser complètement un graphème est l , alors il faudra à DEC $E(l/2) + 2$ passages sur les données pour construire complètement l'ensemble des règles et des scores associés. La seconde concerne la définition du schéma d'expansion du contexte, qui doit être établi *a priori*, et qui s'applique de manière rigide à chaque fois qu'une règle nécessite un supplément de contexte. Il est pourtant clair qu'une plus grande flexibilité serait ici bienvenue, comme on peut le constater sur notre exemple en étudiant les règles extraites pour les lettres *a* et *n* (voir la table B.2).

La prédiction du *a* est (compte tenu de notre modeste lexique) indépendante

1. Formellement, soit R une règle définie par son contexte $C = c_1 \dots c_n$, et sa production $P = \phi$, qui s'applique à N exemples, avec un taux de réussite f . En considérant que la variable aléatoire X définie par $X = (P(c_1 \dots c_n) = \phi)$ suit une loi de Bernouilli de paramètre p , alors l'intervalle de confiance à 95 % de p est $f \pm 1,96 * \sqrt{(f(1-f))/N}$. Ceci n'est valable que pour les règles ambiguës, i.e. où f est différent de 1. Pour les feuilles non-ambiguës, on a utilisé comme borne inférieure la valeur $0,05^{1/f}$. Le score affecté à la règle est la borne inférieure de cet intervalle.

lettre a	lettre n
a → a / # _ ne	n → - / - #
a → ā / # _ ng	n → - / - g
a → ā / # _ n#	n → n / - e
a → ā / # _ n	n → - / -
a → a / - n	
a → ā / -	

TAB. B.2 – Règles extraites pour la prononciation des lettres a et n

de la lettre qui précède, mais dépend des deux lettres qui suivent; le schéma d'expansion du contexte impose toutefois la construction de règles qui intègrent le caractère précédent le a, conduisant à la construction de 6 règles. Par contre, pour désambigüiser le n, il faut examiner et la lettre qui suit et celle qui précède. Le schéma d'expansion du contexte utilisé ici, qui procède précisément à cet examen permet alors d'extraire un ensemble de règles sensiblement plus « naturel » pour le n que pour a.

B.2 Examen de l'algorithme

B.2.1 Apprentissage

Exposition formelle L'intuition qui sous-tend le fonctionnement de l'algorithme étant donnée, il est maintenant possible de le présenter formellement. Pour ce faire, introduisons les notations suivantes. On note $\mathcal{R}$ l'ensemble des règles construites à un moment donné de la procédure, et $\mathcal{R}_a$ les règles qui ont été un moment ambiguës. Si R est une règle de $\mathcal{R}$, on note $C(R)$ le contexte² nécessaire au déclenchement de la règle, et $P(R)$ la production associée; la donnée d'un contexte et d'une production suffit à caractériser complètement une règle. Étant donné une lexie x , on note $c(x_i)$ le contexte courant autour de la lettre i . L'algorithme d'apprentissage de DEC est alors le suivant (voir l'algorithme 9).

Règles ambiguës Afin de compléter notre description de la partie relative à l'apprentissage, il convient de surcroît de noter les deux points suivants. Si le lexique d'apprentissage contient des homographes hétérophones, alors il est impossible de construire une règle fondée uniquement sur l'orthographe pour désambigüiser la phonétique. La manière dont nous avons mis en œuvre DEC permet de tenir compte de ce fait, puisqu'il est possible de créer des règles auxquelles sont associées plusieurs productions différentes. Dans le cas d'homographes hétérophones, ces règles correspondent toujours à des éléments terminaux de l'arbre de décision. Nous avons d'ailleurs étendu cette possibilité aux règles qui correspondent à des nœuds non-terminaux, en modifiant légèrement la procédure de calcul des scores. Étant donné un contexte ambigu (i.e. pour lequel plus d'une production est attestée), il est alors possible de lui associer

2. Dans le sens que nous lui donnons ici, le contexte d'une règle inclut non seulement les conditions de voisinage qu'une lettre doit satisfaire pour que la règle soit déclenchée, mais également la lettre elle-même. Le contexte de la règle: $a \rightarrow \bar{a} / \# _ n\#$ est alors $\#[a]n\#$.

Algorithme 9: DEC - construction de l'arbre

Entrée: Un lexique contenant les formes graphiques et phonétiques alignées; un schéma d'expansion du contexte

Sortie: Un arbre de décision représentant un ensemble de règles dépendant du contexte

DEC-APPRENTISSAGE($\mathcal{L}$, SEC)

```

(1)   $\mathcal{R} \leftarrow \emptyset$ 
(2)   $\mathcal{R}_a \leftarrow \emptyset$ 
(3)  répéter
(4)     $n_a \leftarrow 0$  / *  $n_a$  est le nombre de règles ambiguës */
(5)    pour tout  $x \in \mathcal{L}$ 
(6)      pour  $i = 1$  à  $i = l(x)$ 
(7)         $c(x_i) \leftarrow x_i$ 
(8)        tant que  $\exists R_0 \in \mathcal{R}, c(x_i) = C(R_0)$ 
(9)          étendre  $c(x_i)$ 
(10)       si  $P(R_0) \neq \phi(x_i)$ 
(11)          $\mathcal{R} \leftarrow \mathcal{R} \cup \{R : P(R) = \phi(x_i); C(R) = c(x_i)\}$ 
(12)          $\mathcal{R}_a \leftarrow \mathcal{R}_a \cup \{R_0\}$ 
(13)          $n_a \leftarrow n_a + 1$ 
(14)       sinon si  $R_0 \in \mathcal{R}_a$ 
(15)          $\mathcal{R} \leftarrow \mathcal{R} \cup \{R : P(R) = \phi(x_i); C(R) = c(x_i)\}$ 
(16)       fin
(17)    fin
(18)  fin
(19)  jusqu'à  $n_a = 0$ 
(20)  pour tout  $R \in \mathcal{R}$ 
(21)    calculer  $S(R)$ , le score associé
(22)  fin

```

dans l'arbre de décision plusieurs parties droites différentes, chacune des productions recevant un score fonction de la proportion relative des contextes qui lui correspondent.

De cette manière DEC devient capable de produire plusieurs transcriptions différentes pour une entrée donnée, fonctionnalité qui, nous l'avons vu, est importante pour le traitement spécifique des noms propres. Une réserve toutefois : compte tenu de l'indépendance entre les productions associées aux différentes lettres d'un mot, cette façon de générer des solutions multiples procède sur un mode strictement combinatoire. Reprenons l'exemple précédent (voir table B.1), et supposons que l'on conserve toutes les productions associées aux trois règles R_1 , R_2 et R_3 . Le digraphe *an* sera alors transcrit de quatre manières différentes : /ā/, /an/, /a/ et /ān/. Parmi ces quatre solutions, les deux dernières ne sont jamais attestées, et témoignent des incohérences où peut mener la production *indépendante* des phonèmes associés à des lettres adjacentes.

Conditionnement des productions Nous avons de fait mis en œuvre un moyen de contourner partiellement ce problème, en donnant la possibilité d'inclure dans le contexte d'une règle les phonèmes qui sont disponibles au moment de son application, c'est-à-dire ce qui correspondent à des graphèmes qui pré-

cèdent³ le graphème courant. L'utilisation des phonèmes dans les contextes gauches de règles de production est d'ailleurs une extension classique des systèmes de translittération à base de règles (voir par exemple (Divay 1990)). Par ce biais, on peut d'une part construire des contextes qui permettent de désambigüiser des homographes hétérophones, et d'autre part d'introduire une dépendance entre deux productions successives, ce qui limite singulièrement les phénomènes combinatoires que nous évoquons.

B.2.2 Transcription

La transcription d'un mot inconnu se fait par application de la procédure suivante : pour chaque symbole du mot, on construit l'ensemble des productions possibles, en cherchant dans l'arbre de décision le plus grand contexte possible correspondant au symbole courant. Le fonctionnement est donc identique à celui des systèmes de règles de réécriture habituellement utilisés dans les systèmes de transcription : les règles les plus spécifiques sont appliquées en priorité. Les transcriptions sont ensuite construites en concaténant les phonèmes produits. Si un graphème est ambigu, et correspond à p productions différentes, alors p transcriptions différentes sont proposées. Notons finalement que si DEC doit transcrire une lettre qui n'apparaît pas dans le lexique d'exemple, cette lettre n'est pas transcrite du tout. Dans la pratique, ceci ne se produit que marginalement dès que le nombre d'exemples est suffisamment conséquent.

B.2.3 Note sur la réversibilité

Nous l'avons noté à la section 2.1.2, le principe d'alignement que nous avons utilisé pour créer les données supervisées que requièrent nos programmes autorise l'insertion de nuls graphémiques. Cela signifie qu'un mot tel que *voyage* est présenté au système d'apprentissage sous la forme du couple (voy-age, /vwajaz-/). Dans ce contexte, DEC doit nécessairement apprendre à prédire que certains phonèmes n'ont pas de réel correspondant graphémique. Pour ce faire, lors de l'apprentissage, nous avons simulé entre chaque graphème réel la présence supplémentaire d'un graphème nul, aligné le plus souvent sur un phonème nul. La suite des couples (graphème, phonème) « vue » par DEC pour l'exemple *voyage* est donc : (-, /-/), (v, /v/), (-, /-/), (o, /w/), (-, /-/), (y, /a/), (-, /j/), (a, /a/), (-, /-/), (g, /ʒ/), (-, /-/), (e, /-/), (-, /-/).

Lors de la transcription d'un mot inconnu, le même type de mécanisme est utilisé : avant et après chaque graphème, on simule la présence d'un graphème nul, qui, dans la majeure partie des cas, conduit à la production d'un phonème nul, mais qui, dans des mots tels que *voyage* permet de prédire correctement la présence du /j/ surnuméraire par rapport au groupe *oy*.

3. En parcourant les symboles d'un mot de la droite vers la gauche, on peut en fait aussi bien supposer que le phonème correspondant à la lettre qui *suit* est connu, et donc l'utiliser dans le contexte. Dietterich et Bakiri (1993) utilisent également ce degré de liberté, avec des résultats similaires aux nôtres.

Contexte	Phonèmes		Marqueurs prosodiques	
	% mots	% symboles	% mots	% symboles
<i>fnb</i>	56,670	92,215	46,670	91,223
<i>nfb</i>	56,530	92,135	49,415	91,564
<i>ffbp</i>	54,580	92,031	39,000	89,976
<i>ffb</i>	53,835	91,987	37,545	90,118
<i>fbp</i>	53,730	91,614	40,820	89,523
<i>ffp</i>	53,715	91,828	44,430	89,418
<i>fp</i>	53,430	91,482	42,235	90,083
<i>fb</i>	53,210	91,899	37,595	90,010
<i>bfn</i>	52,815	91,661	45,600	90,898
<i>pf</i>	49,105	90,025	45,625	89,523
<i>bpf</i>	48,990	89,864	40,795	88,499
<i>pbf</i>	48,950	89,877	41,400	88,449
<i>bf</i>	48,490	91,016	37,855	89,874
<i>nb</i>	47,860	90,497	52,380	91,111
<i>bn</i>	44,960	90,071	50,670	90,731
<i>bbfn</i>	44,460	90,037	40,935	89,631
<i>bbf</i>	40,015	89,199	35,670	88,775
<i>bbn</i>	38,050	88,689	44,430	89,418

TAB. B.3 – Performances de DEC en fonction de différentes stratégies d'expansion du contexte. Les données sont celles du lexique NETTALK.

B.3 Quelques résultats

Les résultats que nous présentons dans cette section ont plusieurs visées. Ils vont premièrement constituer des références à l'aune desquelles nos algorithmes analogiques pourront être évalués. Ils vont ensuite nous permettre de continuer à mettre en relief quelques uns des aspects criticables des algorithmes classificatoires.

B.3.1 L'influence du schème d'expansion du contexte

Le tableau B.3 présente les performances moyennes de DEC sur les données du lexique NETTALK (voir l'annexe E), en termes de pourcentage de mots et de symboles correctement transcrits, pour les tâches de prédiction des phonèmes et de prédiction des marqueurs prosodiques. Chaque ligne correspond à une stratégie d'expansion du contexte différente, ces stratégies étant décrites par les lettres *f*, *b*, *n*, *p*, pour désigner respectivement les graphèmes et phonèmes droits et gauches. Ainsi, le schéma *ffb* signifie que le contexte s'étend tant que cela est possible en ajoutant en boucle deux lettres vers l'avant (*forward*) puis une vers l'arrière (*backward*). Appliqué par exemple au symbole *d* du mot *student*, ce schème d'expansion conduit ainsi au contexte maximal: *denut#ts#*.

Les performances qui figurent dans le tableau B.3 sont du même ordre de grandeur que celles rapportées dans (Torkolla 1993) ou (van den Bosch & Daele-

mans 1993)⁴ : dans le meilleur des cas, environ 92 % des phonèmes sont correctement prédits, ce qui correspond à un taux de mots corrects légèrement supérieur à 50 %. La prise en compte des symboles phonétiques dans le contexte se révèle d'assez faible rapport, puisqu'elle permet dans le meilleur des cas un gain d'environ trois points par rapport à la meilleure stratégie purement graphémique (*ffb*); on peut néanmoins penser qu'elle permet d'éviter la production de séquences phonémiques trop improbables. L'importance du choix d'un bon schéma d'expansion du contexte se manifeste par la grande variabilité des performances : plus de 18 points pour les taux de mots corrects séparent en effet le meilleur du moins bon schéma.

Les résultats portant sur la prédiction des marqueurs prosodiques sont sensiblement moins bons. Ceci s'explique à notre avis par la moindre pertinence de l'unité de segmentation considérée : la position de l'accent lexical dépendant fortement de structures métriques de plus haut niveau que le phonème (syllabe, pied), on conçoit bien que fonder sa prédiction sur une analyse graphème par graphème soit relativement peu efficace. Il convient également de noter que le marqueur prosodique à prédire est de nature différente suivant que le graphème correspondant est vocalique, auquel cas le marqueur est alors le niveau d'accentuation de la voyelle; ou consonantique, auquel cas le symbole à prédire dénote l'appartenance du graphème à une attaque ou à une coda de syllabe. Compte-tenu de l'ambivalence de la prédiction à réaliser, il est probable que la contrainte d'utiliser un schéma fixe pour l'expansion du contexte nuit sévèrement aux performances.

On notera qu'utiliser un critère d'information mutuelle permet de guider le choix du SEC. On relève en effet sur le lexique NETTALK les valeurs moyennes suivantes, relatives à la prévision des phonèmes à partir de la seule graphie (voir table B.4, dans laquelle *c* est le caractère courant, *c - 1* le caractère précédent, etc.)

Position	Moyenne	Lettre a	Lettre v	Lettre k
<i>c - 3</i>	0,053	0,058	0,033	0,031
<i>c - 2</i>	0,084	0,117	0,127	0,073
<i>c - 1</i>	0,222	0,326	0,187	0,130
<i>c</i>	1,000	1,000	1,000	1,000
<i>c + 1</i>	0,270	0,259	0,129	0,506
<i>c + 2</i>	0,107	0,077	0,050	0,365
<i>c + 3</i>	0,061	0,025	0,036	0,187

TAB. B.4 – Valeurs moyennes de l'information mutuelle entre graphèmes et phonèmes. Les données sont celles du lexique NETTALK.

Ces valeurs permettent de prédire correctement qu'en moyenne, l'élément le plus important pour transcrire correctement un graphème est le graphème qui le suit. Compte-tenu toutefois de la variabilité entre lettres que laissent entrevoir les informations mutuelles données les lettres *a*, *v* et *k*, on conçoit aisément qu'appliquer le même schème d'expansion à tous les niveaux de l'arbre introduit une rigidité qui nuit très sensiblement aux performances. On notera d'ailleurs que les meilleures performances obtenues pour la prédiction des phonèmes à

4. Des éléments ultérieurs de comparaison sont donnés dans (Yvon 1994b).

rang	correct	prédit	rang	correct	prédit
1	æ	ə	10	ʌ	ə
2	ə	æ	11	e	ɪ
3	ɪ	ə	12	ə	e
4	ə	ɑ: ou ɒ	13	eɪ	ə
5	ə	ɪ	14	ə	ʌ
6	ɑ: ou ɒ	ə	15	əʊ	ə
7	e	ə	16	ɪ	i:
8	ə	eɪ	17	i:	ɪ
9	ɪ	e	18	æ	eɪ

TAB. B.5 – Les erreurs les plus fréquemment commises par DEC. Les valeurs sont calculées sur le lexique NETTALK; le schéma d'expansion du contexte vaut *fnp*.

partir des simples graphèmes utilisent un SEC, *ffb*, qui ne respecte pas tout à fait l'ordre que suggère le critère d'information mutuelle.

B.3.2 Analyse des erreurs

Une étude plus fine permet de mettre à jour la véritable faiblesse de DEC. Nous avons pour cela analysé les erreurs commises sur les 20 000 mots tests en utilisant le SEC le plus le plus favorable. Le tableau B.5 liste les erreurs les plus fréquemment commises. Ces 18 erreurs représentent à elles-seules plus de la moitié des phonèmes incorrectement prédits.

Les erreurs les plus fréquentes portent toutes sur la confusion entre deux voyelles. Il n'y a pas lieu de s'en étonner : une des grandes complexités de la morpho-phonologie de l'anglais (Chomsky & Halle 1968) provient justement de l'instabilité des noyaux vocaliques lors des dérivations morphologiques (que ces dérivations s'accompagnent ou non d'un déplacement de l'accent principal). Ces alternances sont particulièrement ardues à prévoir par une exploration locale du contexte, puisque l'adjonction d'un suffixe à un stemme suffit souvent altérer la qualité de la voyelle initiale. Les exemples de ce phénomène abondent, tels que par exemple les paires *nation-national* (/n'eɪʃən/-/n'æʃənəl/), *vane-vanity* (/v'eɪn/-/v'ænitɪ/), etc.

Le problème que ces erreurs récurrentes révèlent a souvent été décrit comme une incapacité des algorithmes classificatoires à prendre en compte les dépendances à longue distance. À nos yeux, le problème n'est pas tellement celui de la distance, ni celui de la taille de contexte à explorer. DEC, en effet, utilise tout le contexte dont il a besoin : le problème se situe plutôt au niveau de la capacité à distinguer, dans le lexique, les variations qui sont pertinentes de celles qui ne le sont pas. Dans la mesure où le lexique d'apprentissage est de taille finie, il y a tout lieu de penser que les règles de translittération d'un graphème seront d'autant plus difficiles à apprendre qu'elles mettront en jeu des graphèmes plus éloignés, parce que la mise en évidence d'une telle interaction présuppose la capacité d'identifier correctement que les variations orthographiques entre la lettre à transcrire et celle(s) qui le conditionne(nt) sont sans importance pour

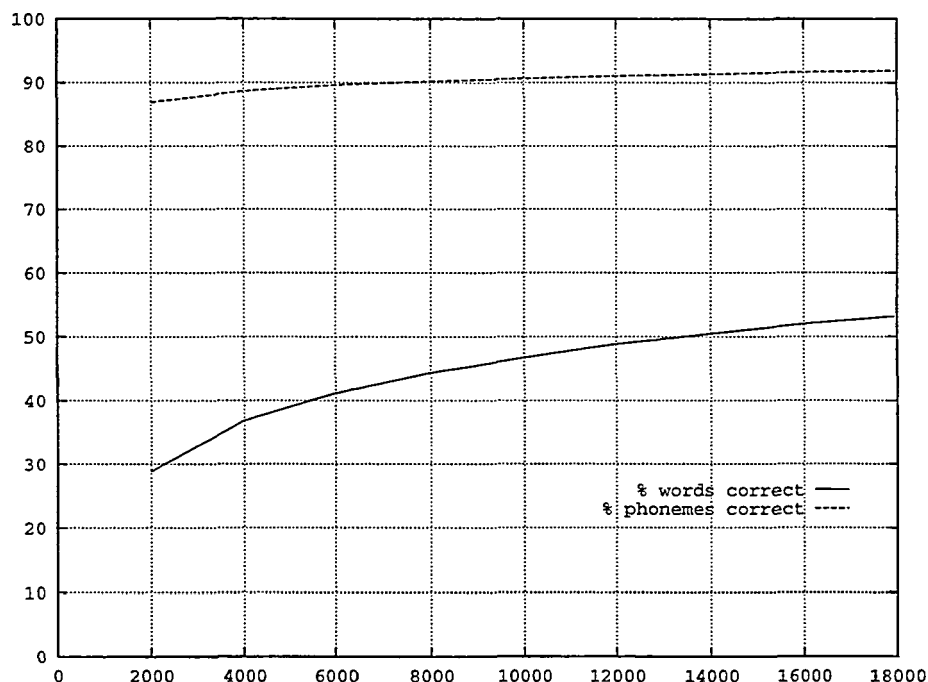


FIG. B.1 – Évaluation de la vitesse d'apprentissage de DEC sur le lexique NET-TALK. Le schéma d'expansion du contexte vaut fb.

la tâche en cours, et ce, sur la base d'indices fréquentiels toujours plus ténus.

B.3.3 Vitesse d'apprentissage et comportement à la limite

Dans ce paragraphe, nous présentons quelques résultats concernant la vitesse d'apprentissage de DEC. La figure B.1 présente les courbes donnant les performances de cet algorithme lorsque l'on fait croître la taille des exemples de 2 000 mots à 18 000.

L'examen de cette courbe permet de constater que DEC est capable de capacités de généralisation intéressantes, puisqu'à partir d'un lexique de 2 000 mots, on obtient déjà un score de prédiction en termes de phonèmes corrects voisin de 87 %. L'allure générale de la courbe est assez trompeuse, en ce qu'elle laisse croire à une augmentation quasi-linéaire des performances avec le nombre d'exemples appris. Il n'en est rien, comme en atteste la courbe reproduite sur la figure B.2.

Sur cette nouvelle courbe, on voit en effet qu'au delà de 50 000 exemples, l'amélioration des performances devient marginale. Nous pensons que ce comportement asymptotique est étroitement lié au caractère essentiellement statistique des généralisations produites par DEC. L'accroissement du volume d'exemples conduit à une nette augmentation du nombre de règles, dont la spécificité réduit toutefois la portée. À l'inverse, de plus en plus de règles sont désambiguïsées sur

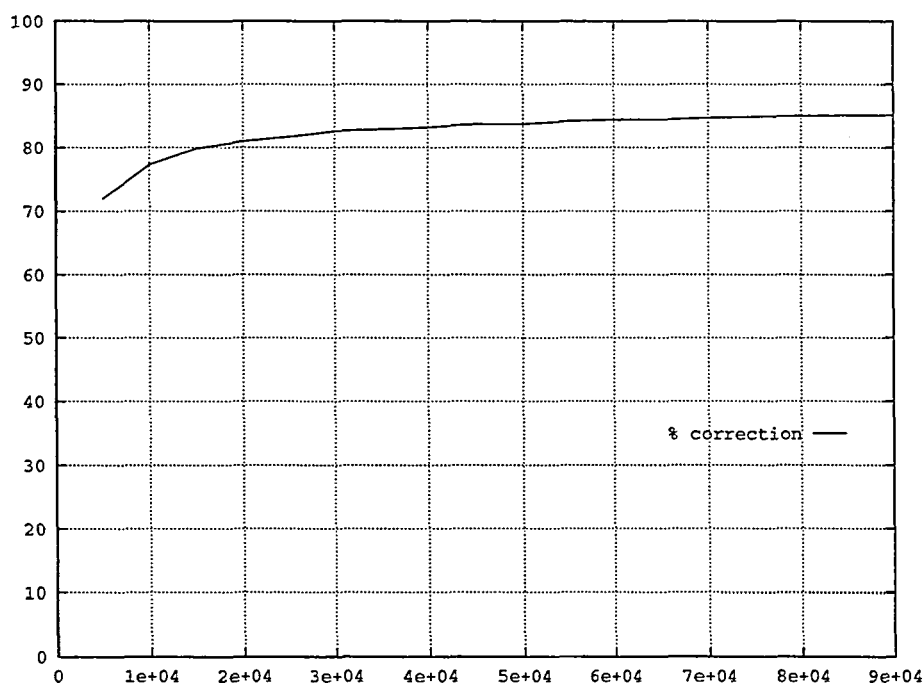


FIG. B.2 – Évaluation de la vitesse d'apprentissage de DEC sur le lexique ENST. Le schéma d'expansion du contexte vaut fb.

des bases statistiques. Il est alors permis de penser que les estimateurs des probabilités d'émissions attachées aux règles sont correctement estimés dès qu'un nombre suffisant d'exemples est vu, nombre au delà duquel ces estimateurs se stabilisent, entraînant une cristallisation du comportement de DEC, et une stabilisation corrélative des performances.

B.3.4 Évaluations complémentaires

Dans la mesure où l'anglais est une langue dont la transcription phonétique à partir de l'orthographe est notoirement complexe, nous avons tenu à explorer les performances de DEC sur des lexiques d'autres langues. Les résultats (voir table B.6 qui donne les meilleurs scores obtenus, ainsi que les schémas d'expansion du contexte correspondants), qui sont directement fonction de la nature plus ou moins phonémique de l'orthographe d'une langue, montrent que cet algorithme obtient des performances plus que satisfaisantes pour le français et le néerlandais, et excellentes pour l'italien.

Les résultats plus mitigés obtenus sur le lexique CMU-20 s'expliquent par la plus grande difficulté de la tâche, puisque les transcriptions phonétiques distinguent, dans ce lexique, les voyelles accentuées des voyelles non-accentuées, conduisant à une nette augmentation du nombre de classes différentes à prévoir. Les résultats également insatisfaisants obtenus sur le lexique allemand (TUB) s'expliquent d'une part par la grande hétérogénéité de ce lexique de noms

Corpus	SEC	% mots	% symboles
BDLEX	<i>fb</i>	86,452	97,858
CMU-20	<i>fb</i>	33,829	87,607
ILC-20	<i>fb</i>	96,016	99,436
ENST-20	<i>bf</i>	76,631	95,085
TUB	<i>fb</i>	49,280	92,211
KUN	<i>fb</i>	86,106	97,952

TAB. B.6 – Performances de DEC sur différents lexiques

propres, et par la taille moyenne plus grande des mots d'autre part. Malgré un taux de succès par phonème très voisin de celui obtenu sur le lexique NETTALK, on observe en effet un déficit de près de 10 points en termes de mots corrects.

B.4 Perspectives et conclusions

B.4.1 Améliorer DEC?

Parvenus à une connaissance intime du fonctionnement de DEC, et tout particulièrement de ses faiblesses, il semblerait naturel d'envisager des améliorations et raffinements à l'approche que cet algorithme met en œuvre. Divers aspects du mécanisme de recherche et d'extraction de règles nous semblent en particulier mériter un surcroît d'intérêt :

- **Filtrage/évaluation des règles** Les performances de DEC reposent pour une large part sur la qualité des règles qu'il est à même d'extraire. Cette extraction s'opère par deux mécanismes bien distincts, à savoir d'une part l'extension du contexte requis pour désambiguïser une séquence, et d'autre part une évaluation fréquentielle de la meilleure production associée à un nœud ambigu. Dans les deux cas se pose de manière cruciale la question de la représentativité du lexique d'apprentissage, dans la mesure où il est légitime de se demander si les règles non-ambiguës sont réellement pertinentes, et si les fréquences qui servent à sélectionner les contextes ambigus ont la moindre validité. Pour améliorer cet aspect de la procédure, on pourrait, amélioration peu coûteuse, mettre de côté une partie des exemples, et utiliser ces données dans une étape de validation statistique indépendante, qui permettrait de mieux évaluer la pertinence et qualité des règles, et de disposer d'un critère pour supprimer les plus idiosyncrasiques d'entre elles. On pourrait également mettre à contribution des outils statistiques, par exemple issus des travaux en statistique exacte, plus adaptés aux effectifs très faibles qui caractérisent les règles que DEC extrait.
- **Sélection du SEC** Les performances de DEC sont très sensibles, nous l'avons vu, à la manière dont le contexte est extrait. Ce dont nous aimerions disposer, c'est d'un moyen de calculer *a priori* l'extension du contexte la plus profitable en termes de performances, la recherche heuristique du meilleur arrangement que nous avons utilisé offrant bien peu de garanties d'optimalité. Une autre amélioration, dont on sent bien qu'elle

repose en partie sur la disponibilité d'un critère de comparaison des SEC, serait d'offrir une plus grande flexibilité dans sa définition, en autorisant par exemple qu'il varie d'un graphème à l'autre.

Si nous n'avons finalement pas persévéré dans ces directions, c'est qu'il nous a semblé que le mal était plus profond, et que même « dopé », DEC (et la classe d'algorithmes qu'il représente) n'était pas un bon candidat pour obtenir l'amélioration qualitative importante des résultats qui est indispensable à une mise en œuvre opérationnelle de systèmes d'apprentissage pour la transcription graphème-phonème. C'est vers d'autres algorithmes que nous nous sommes tourné, délaissant pour un temps les possibilités réelles d'amélioration de DEC.

B.4.2 Conclusion

Nous avons présenté dans cette section un algorithme de transcription fondé sur l'extraction itérative de règles de réécriture contextuelles de plus en plus spécifiques. Un grand nombre de résultats obtenus sur des lexiques variés ont également été présentés et commentés.

Ceci nous a d'une part permis de mettre en évidence, sur des exemples précis, le type de problèmes et de difficultés auxquels est confronté tout mécanisme chargé d'apprendre l'appariement d'une chaîne de graphèmes à une chaîne de phonèmes réalisé par les conventions de la prononciation. Ceci nous a également permis de mettre en relief la nature de l'influence d'un certain nombre de paramètres sur les résultats, tels que la forme et la taille du contexte graphémique considéré, la taille du lexique d'exemples, etc; et de mettre le doigt sur certains handicaps spécifiques d'une approche classificatoire de la transcription comme la difficulté de gérer des ambiguïtés, la nécessité de dupliquer les « règles » de classification pour les deux termes d'un digraphe, etc. En d'autres termes, cette section nous aura permis de passer en revue une large série de faits expérimentaux, qui serviront de support à la critique plus générale que nous développons dans la section 2.4.

Par ailleurs, nous avons pu au cours de cette discussion nous familiariser avec les lexiques qui serviront dans les prochains chapitres de base de test à des approches alternatives de la transcription; constater en particulier que la tâche de les transcrire était très inégalement difficile. Les performances auxquelles DEC nous a permis d'arriver constituent à notre sens des points de repères représentatifs de l'état de l'art en matière de transcription, qui, à ce titre, pourront servir à étalonner de façon réaliste d'autres types de solution à ce problème.

Annexe C

Le système PRONOUNCE

Avec le système PRONOUNCE (Dedina & Nusbaum 1991), Dedina et Nusbaum sophistiquent l'algorithme de Byrd et Chodorow (décrit à la section 3.2.1), en proposant une méthode qui permet de tirer profit non seulement de chaînes initiales et finales, mais de tous les segments du mot à prononcer identifiables dans des mots du lexique. Le propos de Dedina et Nusbaum est clairement affiché : il s'agit de fournir une procédure de prononciation fonctionnant uniquement sur la base d'analogies construites dynamiquement, et par là-même, d'étayer les théories cognitives de la lecture élaborées par Glushko (cf. la présentation des modèles cognitifs de la lecture, à la section 1.2).

C.1 Construction du graphe d'analogies

Dans le modèle de Dedina et Nusbaum, comme dans celui de Byrd et Chodorow, la donnée initiale est un lexique phonétique dans lequel les chaînes graphémiques et phonétiques sont alignées. Confrontée à un mot inconnu x , la procédure que décrivent ces auteurs consiste, dans un premier temps, à rechercher dans le stock de mots disponibles toutes les occurrences des segments de x , et à en récupérer les corrélats phonétiques. Le résultat de cette première recherche est donc un ensemble de chaînes de phonèmes. Le second temps de la procédure consiste à mettre bout-à-bout ces segments, sur la base du principe de construction de chaînes au jeu de domino : le segment x peut prolonger le segment y si le symbole terminal de x est le même que le symbole initial de y . Ré-exprimée dans les termes dans lesquels nous avons exposé JUPA, cette procédure construit des prononciations en recombinaison des segments dont le recouvrement est nécessairement de taille 1. Formellement, cette procédure se décrit par la construction d'un graphe orienté et valué dont les caractéristiques sont les suivantes :

- $\mathcal{G}$ contient deux sommets S (start) et E (end), qui ont des ensembles d'arêtes entrantes (respectivement sortantes) vides.

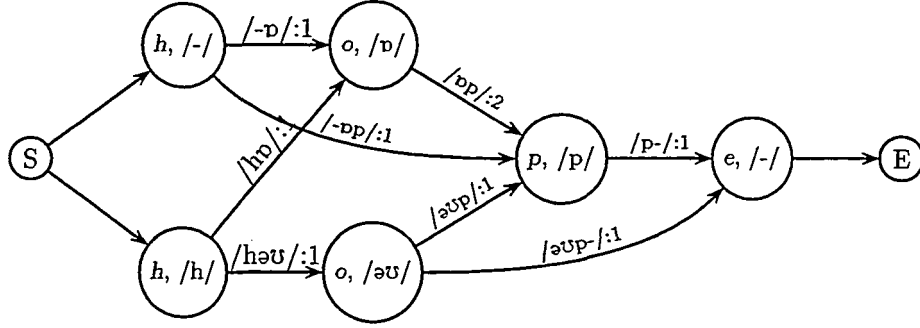


Figure C.1: PRONOUNCE : construction du graphe d'analogies.

- Les autres sommets de $\mathcal{G}$ correspondent à des couples $N = (p, i)$, où p est un phonème et i un indicateur de la position, tels qu'il existe un mot M du lexique et un graphème g tel que g est le i ème caractère du mot inconnu à transcrire, et g est aligné avec p dans M .
- $\mathcal{G}$ contient une arête de $N_a = (p_a, i_a)$ vers $N_b = (p_b, i_b)$ si et seulement si $k = i_a - i_b > 0$, et si il existe un mot M du lexique et deux graphèmes (g_1, g_k) tels que $g_1 \dots g_k$ est une sous-chaîne du mot cible, et $g_1 \dots g_k$ s'aligne sur $p_a \dots p_b$ dans la transcription de M . Les arêtes de $\mathcal{G}$ sont valuées par la fréquence de cet alignement dans le lexique, et sont étiquetées par la chaîne phonétique correspondant à $\phi_M(g_1, g_k)$.
- $\mathcal{G}$ est complété en reliant S à tous les sommets d'indicateur de position égal à 1, et E à tous les sommets d'indicateur de position égal à $l(x)$.

La figure C.1, détaille l'état du graphe $\mathcal{G}$ construit pour inférer la prononciation du mot cible *hope*, lorsque les exemples (*hot*, /hɒt/), (*hose*, /həʊz-/), (*shop*, /ʃɒp/), (*slop*, /slɒp/) et (*slope*, /sləʊp-/), disponibles dans le lexique ont été examinés. Tous les segments (à l'exception du mot complet) qui composent le mot inconnu apparaissent au moins une fois dans le lexique : ainsi *hop* dans *shop*, *ope* dans *slope*, etc. On notera également que la sous-chaîne *op* est identifiée comme appartenant à trois exemples, avec toutefois des prononciations différentes, ce qui conduit à ajouter dans le graphe les arêtes et noeuds correspondants aux suites de phonèmes /ɒp/ et /əʊp/.

C.2 Classement des solutions

Une fois $\mathcal{G}$ construit, chaque chemin dans le graphe permettant de joindre les noeuds initiaux et finaux correspond, par construction, à une prononciation possible du mot inconnu. Ainsi, deux chemins de S à E dans le graphe de la figure C.1 permettent de construire la prononciation (correcte) /həʊp-, un chemin la prononciation (incorrecte) /hɒp-, et deux chemins la prononciation (incorrecte) /-ɒp- pour le mot inconnu *hope*. Les chemins (et les prononciations afférentes) sont ensuite classés par longueur (mesurée par le nombre d'arcs

à parcourir) décroissante et, en cas d'égalité, par poids (mesuré en utilisant les valuations des arcs) croissant. Ce sont donc les deux chemins de taille 4, qui passent respectivement par les noeuds étiquetés $(/h/,h)$, $(/əv/,o)$, $(/əv-/,e)$, et $(/-/,h)$, $(/p/,p)$, $(/-/,e)$ qui conduisent à la meilleure solution. Cette procédure de classement conduit donc à privilégier les solutions qui sont construites par raboutage de fragments longs. Le critère fréquentiel, qui reflète le degré de confiance que l'on peut accorder aux différents segments mis en jeux n'est utilisé que pour départager les *ex æquos*. Ce schème de classement a la propriété (satisfaisante) suivante: si un mot est intégralement contenu dans le lexique, alors sa prononciation sera classée comme la meilleure, puisqu'elle correspond à la construction d'un chemin de taille 3 entre S et E .

La procédure complète est formellement décrite dans l'algorithme 10.

Algorithme 10: Prononciation à base d'analogies - Dedina et Nusbaum

Entrée: Un lexique $\mathcal{L}$ une chaîne $x = c_1 \dots c_{l(x)}$ à transcrire

Sortie: Un ensemble de transcriptions possibles trié de la plus à la moins probable

PRONOUNCE($\mathcal{L}, x$)

- (1) $\phi(x) \leftarrow \lambda$
- (2) Construire itérativement par appariement avec tous les mots du lexique le graphe $\mathcal{G}$
- (3) Construire tous les chemins de S à E dans $\mathcal{G}$ et les classer par ordre de longueur décroissante et de poids croissant.
- (4) Dominoiser les labels des arcs utilisés pour construire les prononciations.
- (5) Proposer les solutions correspondantes par score décroissant.

Dans notre implantation de cet algorithme, nous avons utilisé une procédure de stockage des segments identique à celle mise en œuvre dans la procédure SMPA. Ceci nous a permis, comme pour SMPA, d'introduire deux paramètres supplémentaires, qui sont les tailles minimales et maximales d'un segment utilisable pour la recombinaison. Pour plus de détails sur cette procédure de stockage, on se reportera à l'algorithme 2, présenté à la section 3.3.2.

Les résultats de Dedina et Nusbaum

La procédure de tests mise en place par Dedina et Nusbaum vise simplement à appliquer et à tester les hypothèses de Glushko (1979) concernant la possibilité de prononcer uniquement sur la base d'analogies. Dans cette optique, seuls les 70 pseudo-mots utilisés dans les expériences originales de Glushko ont été transcrits et évalués. Partant d'un lexique d'environ 20 000 mots, PRONOUNCE parvient dans 91 % des cas à fournir une prononciation correcte (i.e. qui soit également proposée par au moins un sujet dans le test original). À titre de comparaison, Dedina et Nusbaum indiquent que le système commercial DECtalk obtient pour la même tâche un taux de succès de l'ordre de 97 %. Il est difficile de comparer ces résultats avec ceux présentés par ailleurs, dans la mesure où les pseudo-mots testés sont particulièrement simples, puisqu'il ont été construits

Contrainte	% mots	% silence	% symboles
$2 \leq x \leq 7$	56,760	0,265	91,856
$2 \leq x \leq 8$	56,680	0,310	91,782
$2 \leq x \leq 258$	56,565	0,325	91,686
$2 \leq x \leq 6$	56,465	0,275	93,255
$2 \leq x \leq 5$	52,535	0,235	91,016
$3 \leq x \leq 7$	52,045	2,510	89,076
$3 \leq x \leq 8$	51,990	2,505	88,956
$3 \leq x \leq 6$	51,500	2,495	88,924
$3 \leq x \leq 5$	46,625	2,485	87,790
$2 \leq x \leq 4$	42,450	0,235	88,766
$3 \leq x \leq 4$	37,845	2,655	85,214
$2 \leq x \leq 3$	24,710	0,260	83,318
$3 \leq x \leq 3$	12,525	4,715	73,957

TAB. C.1 – Performances de PRONOUNCE en fonction de la taille des segments. Les données sont celles de NETTALK

par remplacement de la première lettre de mots mono-syllabiques existant dans le lexique. Ce biais de sélection ne permet pas non plus de se faire une idée du taux d'échec de la procédure, i.e. de la proportion de mots pour lesquels il n'existe pas dans $\mathcal{G}$ de chemin conduisant de S à E , et pour lesquels aucune transcription ne peut être produite.

C.3 Les performances réelles de PRONOUNCE

Le premier résultat important à mentionner est la très faible proportion des mots imprononçables, et ce même quand le nombre d'exemples est faible. En effet, les expériences avec le lexique NETTALK révèlent qu'après apprentissage de 2 000 mots, le taux de silence de PRONOUNCE est de l'ordre de 4 %; et que ce pourcentage décroît très vite, puisqu'au-delà de 6 000 exemples il tombe en dessous de 1 %. Au final, en utilisant une base d'exemples complète (18 000 mots), le taux de silence est de l'ordre de trois mots pour 1 000. Des résultats similaires, présentant des taux de silence marginaux, ont été observés sur tous nos lexiques. L'utilisation de segments internes, et non plus simplement de segments initiaux ou finaux (comme c'est le cas de la procédure JUPA, présentée à la section 3.2), est donc une solution efficace pour faire disparaître les problèmes liés à la non-prononciation de certains mots. Ceci justifie donc largement *a posteriori* que Dedina et Nusbaum ne se soient pas préoccupés des problèmes de silence, qui apparaît être un phénomène marginal.

Cependant, si PRONOUNCE parvient à donner une transcription de presque n'importe quelle chaîne, ces transcriptions ne sont pas, loin s'en faut, toujours correctes, comme en témoigne la table C.1, où figurent les résultats de diverses expériences avec le lexique NETTALK en fonction des contraintes de taille imposées aux segments extraits.

Corpus	Contrainte	% mots	% symboles
NETSTRESS	$2 \leq x \leq 258$	48,075	90,743
BDLEX	$2 \leq x \leq 8$	82,971	96,647
CMU-20	$2 \leq x \leq 8$	35,776	86,680
ILC-20	$2 \leq x \leq 7$	93,068	98,822
ENST-20	$2 \leq x \leq 6$	74,250	94,377
TUB	$2 \leq x \leq 8$	37,627	85,452
KUN	$2 \leq x \leq 6$	81,663	96,587

TAB. C.2 – Performances de PRONOUNCE sur différents lexiques.

Les résultats reproduits dans ce tableau forcent donc à relativiser le succès de cette approche à base d'analogie. Dans le meilleur des cas, seuls un peu plus d'un mot sur deux sont correctement transcrits. Les performances sont ici très proches de celles de DEC. On notera toutefois un élément intéressant, à savoir que la stratégie qui consiste à conserver tous les fragments n'est pas celle qui conduit aux meilleures performances. En effet, c'est en ne conservant que des chaînes de taille inférieure ou égale à sept caractères que nous avons obtenu les meilleurs résultats sur le lexique NETTALK.

Des résultats analogues, systématiquement comparables à ceux de DEC, ont été obtenus sur les différents lexiques que nous avons traité (voir la table C.2).

C.4 Conclusion

Nous avons, dans ce chapitre, présenté et évalué une version légèrement étendue de l'algorithme de Dedina et Nusbaum. Le principal enseignement que nous pouvons tirer de nos expériences avec cet algorithme est que ses performances sur des lexiques de tests de taille importante sont sensiblement inférieures aux résultats obtenus sur un corpus de pseudo-mots dans le cadre de la mise en application des principes cognitifs identifiés par Glushko.

Un second enseignement de ces expériences est qu'apparemment, le principe de transcription mis en œuvre dans cet algorithme est suffisamment peu contraignant pour que pratiquement tous les mots inconnus soient effectivement d'être transcrits. Ce résultat était absent de l'article original de Dedina et Nusbaum.

En dernier lieu, il ressort de ces expériences que l'analogie, telle qu'il est possible de l'appréhender dans une telle approche, reste relativement mal définie. Il est en effet manifeste que le mécanisme de transcription contient en lui-même, compte-tenu d'un lexique d'exemples « bien choisi », la capacité de produire des transcriptions qui sont identiques à celles que peut produire un humain dans des conditions expérimentales simulant une tâche de lecture, et qui peuvent par là-même être qualifiés d'analogiques. Dans le cas général toutefois, dresser la liste des analogues desquels le système de transcription s'inspire est loin d'être évident, puisqu'un grand nombre d'exemples peuvent contribuer à une prononciation donnée. Il est plus juste, à notre sens, de dire que cet algorithme n'extrait pas des analogies, mais qu'il en produit, c'est-à-dire qu'il élabore des transcriptions, qui, sous certaines conditions, peuvent être inter-

prétées comme résultant de l'application d'un principe relativement vague de ressemblance entre les chaînes graphiques.

Annexe D

Protocole général d'évaluation des performances

D.1 Mesure des performances

Les performances des différents algorithmes ont été mesurées en respectant le protocole expérimental suivant. Pour chacun des lexique dont nous disposons, une procédure de sélection aléatoire itérée dix fois a permis la construction de 10 paires (ensemble d'apprentissage, ensemble de test). Pour chaque jeu de paramètre de chacun des algorithmes présentés, dix expériences ont donc été menées, dont le résultat moyen est utilisé comme score final. Lors de chacune de ces expériences, les mesures suivantes ont été réalisées :

- *correction par mot* : c'est le rapport entre le nombre total de mots correctement transcrits (i.e. dont la transcription produite est exactement¹ la même que celle répertoriée dans notre lexique) et le nombre total de mots à transcrire. De nombreux auteurs ont critiqué la validité de ce score, au titre de sa sévérité. Ce calcul ne tient en effet nul compte de la finalité pour laquelle les systèmes de transcription ont été développés, à savoir la synthèse vocale. Compte-tenu d'une part de la qualité propre des systèmes de synthèse, et d'autre part de la relative variabilité naturelle autorisée par les processus cognitifs mis en œuvre au moment de la reconnaissance, on peut à bon droit arguer que certaines erreurs sont moins graves que d'autres, et qu'en particulier il en existe qui ne nuisent pas à la reconnais-

1. Signalons toutefois le cas particulier des évaluations du lexique BDLEX. Dans ce lexique (voir annexe F), un certain nombre d'archiphonèmes sont utilisés afin de prendre en compte des phénomènes de variabilité. Nous avons ignoré les différences entre nos transcriptions et celles disponibles dans la base de données à chaque fois que le symbole prédit était *plus spécifique* que celui de la transcription de référence. Ceci nous a conduit à considérer que prédire un /ɔ/ quand le symbole de référence laissait les deux options (/ɔ/ et /o/) possibles ne comptait pas pour une erreur.

sance. Ce score a donc tendance à sous-estimer les performances réelles des systèmes. Ceci a d'ailleurs été confirmé par de nombreuses comparaisons avec des évaluations alternatives, notamment des tests auditifs fondés sur la synthèse des transcriptions produites. On notera par contre que l'argument qui consiste à considérer directement le nombre de phonèmes en erreur comme un indicateur absolu de la qualité d'une transcription nous semble tout particulièrement spécieux : il existe nous semble-t-il un certain nombre d'exemples depuis longtemps discutés par les linguistes pour lesquels la confusion sur un phonème entraîne une confusion sur le mot ; c'est d'ailleurs sur la base de ces exemples (et sur les paires minimales qu'ils constituaient) qu'historiquement la notion de phonèmes a été introduite.

Compte-tenu finalement de la simplicité de mise en œuvre de cette mesure par rapport à la construction de tests d'écoutes, c'est elle que nous avons utilisé pour évaluer les performances des divers algorithmes.

- *silence* : c'est le rapport du nombre total de mots pour lesquels le système testé échoue à produire une transcription ou produit une transcription vide et le nombre total de mots à transcrire.
- *justesse par mot* : c'est le rapport entre le nombre total de mots correctement transcrits et le nombre de mots pour lesquels le système émet une prédiction (c'est-à-dire n'est pas silencieux).
- *correction par phonème* : c'est le rapport entre le nombre total de phonèmes (non « nuls ») correctement prédits et le nombre total de phonèmes (non « nuls ») dans le lexique de test. Ce score est calculé sur la base de la distance lexicale d entre la chaîne de phonèmes de référence et la chaîne effectivement produite. Cette distance mesure le nombre d'opérations élémentaires (insertion, suppression, échange) nécessaires pour transformer la chaîne de référence en la chaîne produite, chaque opération augmentant la distance d'une unité. Si x dénote alors la transcription correcte, et y la valeur prédite par le système testé, alors on considère que le nombre de symboles corrects vaut : $l(x) - d(x, y)$. Notons qu'une version « améliorée » du calcul de la distance lexicale a été utilisée, qui permet non seulement de mesurer ladite distance, mais fournit également la liste des transformations élémentaires correspondantes. Ceci nous a également permis d'établir la liste phonème par phonème des erreurs les plus fréquemment commises par tel ou tel algorithme.
- *correction par phonème* : c'est le rapport entre le nombre total de phonèmes (non « nuls ») correctement prédits et le nombre total de phonèmes (non « nuls ») dans l'ensemble de mots du lexique de test pour lesquels le système testé est en mesure de faire une prédiction.

D.2 Comparaison de deux algorithmes

Pour comparer les résultats de deux algorithmes, on considère pour chacun des algorithmes la variable aléatoire X correspondant au pourcentage de mots correctement prédits pendant une expérience. On dispose de 10 réalisations de

cette variable aléatoire, pour laquelle il est permis, compte-tenu de la taille des populations sur lesquelles elle est calculée, de faire l'hypothèse de normalité. Ceci autorise alors de mettre en œuvre un test bilatéral de comparaison des moyennes sur échantillon appariés. Sous l'hypothèse nulle d'égalité des moyennes, $t_{n-1} = \sqrt{(n-1)} | (X_1 - X_2) | / s_d$, où n est le nombre d'expériences, et s_d la variance empirique de la différence, suit une loi de student de paramètre $n - 1$. Pour $n = 10$, si la valeur observée de t_{n-1} excède respectivement les valeurs 2,262, 3,250, et 4,781, alors on peut considérer que la différence entre les scores est significative aux seuils de 0,05, 0,01 et 0,001.

Annexe E

Le lexique NETTALK

Le lexique NETTALK est le lexique qui a été utilisé dans les expériences originales de Sejnowski et Rosenberg (1987). Ce lexique contient 20,008 mots anglais ainsi qu'une transcription phonétique *manuellement alignée*. Cette transcription contient des informations sur la nature des segments, mais encode également trois niveaux distincts d'accentuation pour les voyelles, ainsi que des informations sur la position métrique des consonnes. Chaque entrée de ce lexique est décrite par une succession de champs, chaque graphème du mot correspondant à une ligne. Par exemple, le mot *tradition* est décrit par la série de succession de lignes suivante :

```
<>
t => t, >;
r => r, >;
a => x, 0;
d => d, >;
i => I, 1;
t => S, <;
i => -, 0;
o => x, <;
n => n, <;
<>
```

Pour chaque graphème de ce mot, le premier symbole à droite du signe => décrit le phonème (voir la table E.1) associé dans un schème d'alignement bi-univoque. Le symbole — est utilisé pour marquer un phonème « nul ». Le second symbole indique, pour une voyelle, le niveau d'accent (0, 1 ou 2); pour une consonne, il précise si la consonne appartient à l'attaque syllabique (>), ou à la rime (<). Ainsi, l'extrait précédent de ce lexique permet d'affirmer que l'accent principal de *tradition* porte sur la voyelle /I/.

Nous avons utilisé ce lexique pour deux tâches bien distinctes : la prédiction des phonèmes, et la prédictions des marqueurs supra-segmentaux. Dans ces deux cas, nous avons isolé les séquences de symboles correspondant à la tâche à évaluer, et éliminé les possibles entrées dupliquées des fichiers ainsi créés (en

particulier, certains mots ne se distinguent que par la position de l'accent lexical: si cette information est ignorée, les deux entrées correspondantes seront identiques).

Le lexique correspondant aux données de la tâche de transcription a été baptisé NETTALK, et contient 19 940 entrées distinctes. L'entrée correspondante au mot *tradition* dans ce fichier est:

tradition trxdIS-xn

Le lexique correspondant aux données de la tâche de prédiction des marqueurs supra-segmentaux a été baptisé NETSTRESS, et contient 19 9971 entrées distinctes. L'entrée correspondante au mot *tradition* dans ce fichier est:

tradition >>0>1>0<<

L'alphabet phonétique utilisé pour décrire les chaînes segmentales est reproduit à la table E.1. Il est à noter que cet alphabet contient des symboles spéciaux pour décrire les consonnes « vocaliques »; et utilise également des symboles spéciaux pour noter certains diphtonges (/ks/, /gz/, etc.) qui posent des problèmes d'alignement. En particulier, l'utilisation de symboles atomiques pour noter les diphtonges permet de traiter les configurations dans lesquelles une entrée graphique est plus courte que l'entrée phonétique correspondante.

NETTALK	IPA	Comme dans ...	Transcrit ...
[i]	[i]	bead, see	bid, si
[a]	[a]	bard, calm	bard, kam
[c]	[ɔ]	born, cork	bcrn, kcrk
[u]	[u]	boon, fool	bun, ful
[R]	[ɜ]	burn, work	bRn, wRk
[I]	[ɪ]	sit, pity	sIt, pItY
[E]	[e]	set, less	sEt, lEs
[@]	[æ]	sat, apple	s@t, @pL
[^]	[ʌ]	fun, come	f^n, k^m
[a]	[ɒ]	fond, wash	fand, waS
[U]	[ʊ]	full, soot	fUL, sUt
[x]	[ə]	composer, above	kxmpczR, xb^v
[e]	[ei]	bay, fate	be, fet
[A]	[ai]	buy, lie	bA, lA
[O]	[oi]	voice, boy	vOs, bO
[o]	[əʊ]	no, ago	no, xgo
[W]	[aʊ]	now, plough	nW, plW
[I]	[Iə]	tier, bier	tIr, bIr
[@]	[eə]	tare, fair	t@r, f@r
[L]	[əl]	typical	tIpIkL
[M]	[əm]	truism, sadism	truIzM, sedIzM
[N]	[ən]	listen, fountain	lIsN, fWntN
[Y]	[yʊ]	duo, curlew	dYo, kRIY
[y]	[j]	yet, million	yEt, mIllyxn
[w]	[w]	wet, bewail	wEt, bIwel
[]	[w]	why, while	*A, *Al
[h]	[h]	hat, preheat	h@t, prihit
[p]	[p]	pat, pope	p@t, pop
[t]	[t]	tab, strut	t@b, str^t
[k]	[k]	cot, kiss	kat, kIs
[b]	[b]	bat, baby	b@t, bebi
[d]	[d]	dab, amend	d@b, xmEnd
[g]	[g]	got, agog	gat, xgag
[f]	[f]	fine, raffle	fAn, r@fL
[s]	[s]	sit, rice	sIt, rAs
[S]	[ʃ]	ship, sugar	SIp, SUGR
[T]	[θ]	thin, math	TIn, m@T
[C]	[tʃ]	chance	C@ns
[v]	[v]	vine, river	vAn, rIvR
[z]	[z]	Pods, buzz	padz, b^z
[Z]	[ɜ]	measure	mEZR
[D]	[ð]	this, other	DIs, ^DR
[J]	[dʒ]	just, edge	J^st, EJ
[l]	[l]	little, place	lItL, ples
[r]	[r]	ran, steer	r^n, stIr
[m]	[m]	ram, mammoth	r@m, m@mxT
[n]	[n]	ran, nut	r^n, n^t
[G]	[ŋ]	rang, bank	r@G, b@nG
[!]	[ts]	nazi, pizza	na!i, pi!x
[X]	[ks]	convexity	kxnvEXxti
[K]	[kʃ]	anxious	@GKxs
[#]	[gz]	exit	E#xt

TAB. E.1 – Alphabet phonétique utilisé dans le lexique NETTALK

Prononciation par analogie : motivations, formalisations, évaluations

Annexe F

Le lexique BDLEX

Le lexique BDLEX (Pérennou *et al.* 1991) est extrait d'un lexique du Français contenant environ 23000 mots et mots composés. La version de ce lexique que nous avons utilisé est la version 1.2 du lexique général étendu. Pour chaque entrée lexicale, plusieurs informations sont disponibles : catégorie(s) morpho-syntaxique(s), type de flexion, dérivations possibles, transcription phonétique La transcription phonétique dans BDLEX est relativement complexe; aussi quelques précisions supplémentaires sont-elles requises pour décrire exactement ce qui en a été fait.

Le codage de la transcription phonétique occupe trois champs de la base de données. Le premier contient une représentation phonologique assez abstraite, qui inclut les frontières de syllabes et de pied métrique. Le second champ décrit le comportement phonologique de la finale du mot : en particulier, il indique pour une consonne finale si c'est une consonne latente ou optionnelle, et dans ce dernier cas, quelle est la forme préférée, celle avec ou sans la consonne. Le troisième champ contient une décomposition du mot sous la forme de *phonogrammes*, qui représentent en fait l'alignement entre lettres (ou groupes de lettres) et phonèmes (ou groupes de phonèmes). Cette décomposition phonogrammatique, comme le note Catach (1984), est un bon outil pour déceler automatiquement les zones d'un mot qui risquent de favoriser les fautes d'orthographe. Ce sont celles pour lesquelles un groupe de lettre est associé à un groupe phonématique pour lequel de nombreuses formes orthographiques sont attestées dans le lexique.

Le tableau F.1 donne quelques exemples d'entrées lexicales de BDLEX (seuls les champs relatifs à la transcription phonétique ont été conservés). La table de correspondance entre le codage des symboles phonétiques et le codage IPA est donnée dans la table F.3.

Ainsi la transcription du mot *ananas* comporte un /s/ final optionnel, dont la réalisation est usuelle; le mot *divin* comporte un /n/ final latent : en cas de liaison, *divin* est alors réalisé /divi#n/.

À partir des deux premiers champs, nous avons reconstitué la transcription phonétique du mot lorsqu'il est isolé. Pour ce faire nous avons appliqué la pro-

Entrée lexicale	Transcription	Finale	Phonogrammes
ananas	A/nA/nA	s 1	(a/A)(n/n)(a/A)(n/n)(a/A)(s/s.)
divin	/di/vi	n"	(d/d)(i/i)(v/v)(in/in")
grand	/gra [^]	d"	(g/g)(r/r)(an/a [^])(d/d")
voyage	/vwA/jAJ	e	(v/v)(oy/wAj)(a/A)(g/J)(e/e)
yaourt	/*jA/ur	t 0	(/*)(y/j)(a/A)(ou/u)(r/r)(t/t.)

TAB. F.1 – Quelques exemples tirés de BDLEX

cédure suivante :

- nasalisation de toutes les voyelles qui étaient suivies d'un /n/ final latent.
- suppression de toutes les consonnes latentes finales
- suppression des consonnes finales optionnelles lorsqu'elles ne se réalisent pas dans la prononciation préférée du mot.

Le résultat de ces différents ajustements conduit, à partir de l'exemple précédent, au lexique suivant (voir table F.2) :

Entrée lexicale	Transcription	Phonogramme
ananas	A/nA/nA	(a/A)(n/n)(a/A)(n/n)(a/A)
divin	/di/vi [^]	(d/d)(i/i)(v/v)(in/i [^])
grand	/gra [^]	(g/g)(r/r)(an/a [^])(d/)
voyage	/vwA/jAJ/Je	(v/v)(oy/wAj)(a/A)(g/J)(e/e)
yaourt	/*jA/urt	(y/*j)(a/A)(ou/u)(r/r)(t/t)

TAB. F.2 – Quelques exemples tirés de BDLEX

Enfin, nous avons mis à jour les phonogrammes de façon à supprimer les configurations dans lesquelles un graphème était aligné sur un phonème nul. Ces configurations étaient majoritairement la conséquence de la suppression de consonnes finales latentes. Ainsi le phonogramme du mot *grand*, transformé durant cette phase de [(g/g)(r/r)(an/a[^])(d/)] à [(g/g)(r/r)(and/a[^])]. Ceci nous a permis de disposer, pour chaque entrée, d'un alignement multivoque de bonne qualité. Pour la plupart de nos expériences, nous avons déduit de cet alignement multivoque un alignement univoque, en complétant par des symboles nuls les phonogrammes pour lesquels les séquences graphiques et phonétiques n'étaient pas de même longueur. Ceci nous permit par exemple de reconstruire les alignements univoques suivants : (*divin*, /divi[^]-/), (*voy-age*, /vwAjaJe/), etc.

À l'issue de cette troisième étape de traitement, nous avons construit deux corpus pour nos évaluations. Le premier, BDLEX1 ne contient pas d'information sur les catégories morpho-syntaxiques de ses entrées. Les entrées identiques (correspondant aux mots ne différenciant que par leur catégorie morpho-syntaxique) ont été fusionnées, pour aboutir à un lexique final de 22 553 entrées.

Dans le second, BDLEX2, les catégories morpho-syntaxiques principales¹ associées aux entrées ont été conservées. Le nombre d'entrées de ce lexique, après élimination des enregistrements dupliqués, est de 22 824.

Une dernière remarque concernant les transcriptions phonétiques : dans BDLEX, les voyelles à deux timbres ne sont distinguées que dans les cas où la distinction est obligatoire *dans tous les dialectes du français*. Dans les autres cas, pour lesquels cette distinction est susceptible d'être neutralisée (neutralisation qui a essentiellement lieu dans les parlers méridionaux), des archi-phonèmes ont été utilisés dans les transcriptions. Ceci explique pourquoi l'alphabet phonétique utilisé dans BDLEX contient quatre symboles ajoutés à la liste habituelle des phonèmes du français. Le nombre total de symboles de cet alphabet est alors de 48 (voir la table F.3).

1. BDLEX utilise les catégories morpho-syntaxiques principales suivantes : adverbe, conjonction, déterminant, adjectif ou nom féminin, adjectif ou nom (masculin ou féminin) adjectif, interjection, adjectif ou nom masculin, nom, préposition, pronom, verbe à l'infinitif

bdlex	API	As in ...	Transcribed as ...
[i]	[i]	il, vie, lyres	i l, v i, l i r
[j]	[e]	blé, jouer	b l }, J u }
[f]	[ɛ]	laine, jouet	l { n e
[&]	[e] or [ɛ]	lait, jouet	l &, J u &
[a]	[a]		
[@]	[a]		
[A]	[a] or [ɑ]	plat, patte	p l A, p A t
[o<]	[ɔ]	mort, donner	m o< r, d o< n }
[o>]	[o]	dôme, beau	d o> m, b o>
[o]	[ɔ] or [o]	daurade	d o r A d e
[u]	[u]	genou, roue	J e n u, r u
[y]	[y]	rue, venue	r y, v e n y
[EU]	[ø]	peu, deux	p EU, d EU
[OE]	[œ]	peur, meuble	p OE r, m OE b l
[E]	[œ] or [ø]	peuplé, bleuet	p E p l }, b l E &
[e]	[ə]	le, premier	l e, p r e m j e
[i^]	[ē]	matin, plein	m A t i^, p l i^
[a^]	[ā]	sans, vent	s a^, v a^
[o^]	[ō]	ombre, bon	o^ b r e, b o^
[y^]	[œ]	lundi, brun	l y^ d i, b r y^
[j]	[j]	yeux, pied	y EU, p j }
[w]	[w]	étoile, chouette	} t w A l e, \$ w } t e
[!]	[ɥ]	huile, lui	! i l e, l ! i
[p]	[p]	père	p { r e
[t]	[t]	terre	t { r e
[k]	[k]	cou, sac	k u, s A k
[b]	[b]	bon, robe	b o^, r o< b e
[d]	[d]	dans, aide	d a^, { d e
[g]	[g]	gare, bague	g A r e, b A g e
[f]	[f]	feu, phare	f EU, f A r e
[s]	[s]	sale, tasse	s A l e, t A s e
[sh]	[ʃ]	chat, tache	\$ A, t A \$ e
[v]	[v]	vous, rêve	v u, r { v e
[z]	[z]	zéro, maison	z } r o>, m & z o^
[j]	[ʒ]	je, gilet	J e, J i l &
[l]	[l]	lent, sol	l a^, s o< l
[r]	[ʀ]	rue, venir	r y, v e n i r
[m]	[m]	main, femme	m i^, f A m e
[n]	[n]	nous, tonne	n u, t o< n e
[g^]	[ɲ]	agneau, vigne	a g^ o>, v i g^ e
[*]	[h]	hop	* o< p
[ng]	[ŋ]	camping	c a^ p i n g

TAB. F.3 – *Alphabet phonétique utilisé dans le lexique BDLEX*

Annexe G

Les lexiques d'ONOMASTICA

Nous avons utilisé pour nos évaluations quatre lexiques de noms propres (plus précisément de patronymes) développés dans le cadre du projet européen ONOMASTICA (Schmidt *et al.* 1993). Ces lexiques présentent des caractéristiques légèrement différentes à la fois en termes de leur taille, mais également en termes de leur contenu.

Dans tous les cas, nous avons adapté les transcriptions originales, en faisant disparaître les frontières de syllabe, et les accents lexicaux. Par contre, nous avons, dans les cas où cela était pertinent, distingué plusieurs niveaux de longueur, aussi bien pour les voyelles que pour les consonnes.

G.1 Le lexique enst

Le lexique ENST a été développé sous notre direction à l'École Nationale Supérieure des Télécommunications de Paris, avec la collaboration d'Axelle Culubret et de Monica Margulis.

Ce lexique contient environ 98 000 patronymes, pour une part sélectionné en partie sur des critères fréquentiels (en particulier il contient intégralement les 25 000 patronymes les plus fréquents), et en partie par des tirages aléatoires. À ce titre, il est relativement révélateur de la diversité linguistique du stock de patronymes français. Toutes les transcriptions de ce lexique ont été contrôlées manuellement. Une des difficultés que pose ce lexique provient du nombre relativement important de prononciations multiples. En particulier, pour tous les mécanismes qui sont capables d'apprendre « par cœur », une proportion relativement négligeable des erreurs observées provient de ce qu'une prononciation est disponible dans le lexique d'exemple, et qu'une prononciation alternative est dans le jeu test. Dans ce cas, la prononciation apprise est restituée, et est jugée incorrecte.

Nous avons été amené à utiliser, dans certaines expériences, un sous-ensemble

plus restreint de ce lexique, contenant les 20 000 patronymes les plus fréquents. Ce lexique est utilisé sous le nom de ENST-20.

La table G.1 contient l'alphabet phonétique utilisé pour décrire les prononciations contenues dans ce lexique.

G.2 Le lexique Ilc

Le lexique ILC est un lexique de noms propres italiens, construit pendant le projet ONOMASTICA, sous la direction de Vito Pirrelli à l'Institut d'Informatique Linguistique (ILC) de Pise. Ce lexique a été gracieusement mis à notre disposition par V. Pirrelli.

Ce lexique contient environ 90 000 noms propres, dont la transcription a été validée manuellement par des experts phonéticiens.

Nous avons été amené à utiliser, dans certaines expériences, un sous-ensemble plus restreint de ce lexique, contenant 20 000 patronymes choisis aléatoirement. Ce sous-lexique est utilisé sous le nom de ILC-20.

Les transcriptions de ce lexique sont décrites par un alphabet de 60 phonèmes.

G.3 Le lexique Kun

Le lexique KUN est un lexique de noms propres néerlandais, construit pendant le projet ONOMASTICA, par Emmy Konst et Henk van Heuvel, sous la direction de Lou Boves, à l'Université Catholique de Nimègue (KUN). Ce lexique a été gracieusement mis à notre disposition par H. van Heuvel.

Ce lexique contient environ 25 000 patronymes, dont la transcription a été validée manuellement par des experts phonéticiens, et qui ont été tirés au hasard dans un ensemble d'environ 80 000 noms. Ce lexique présente la particularité d'être très homogène, dans la mesure où les noms d'origine étrangère, qui étaient précisément identifiés dans la base de donnée originale, ont été mis de côté pour les tests.

Les transcriptions de ce lexique sont décrites par un alphabet de 46 phonèmes.

G.4 Le lexique Tub

Le lexique TUB est un lexique de noms propres allemands construit pendant le projet ONOMASTICA, par Andreas Mengel, sous la direction de Klaus Feldbaum, à l'Université Technique de Berlin (TUB). Ce lexique a été gracieusement mis à notre disposition par A. Mengel.

Ce lexique contient environ 25 000 patronymes, sélectionné aléatoirement

ENST	IPA	Comme dans ...	Transcrit ...
[i]	[i]	il, vie, lyres	i l, v i, l i r
[e]	[e]	blé, jouer	b l e, j w e
[ai]	[ɛ]	lait, jouet	l a i, j w a i
[a]	[a]	plat, patte	p l a, p a t
[aa]	[ɑ]	bas, pâte	b a a, p a a t
[o]	[ɔ]	mort, donner	m o r, d o n e
[oo]	[o]	dôme, beau	d o o m, b o o
[ou]	[u]	genou, roue	j @ n o u, r o u
[u]	[y]	rue, vêtu	r u, v a i t u
[eu]	[ø]	peu, deux	p e u, d e u
[oe]	[œ]	peur, meuble	p o e r, m o e b l
[@]	[ə]	le, premier	l @, p r @ m y e
[@@]	[ə]	matelot	m a t @ @ l o o
[in]	[ɛ̃]	matin, plein	m a t i n, p l i n
[an]	[ɑ̃]	sans, vent	s a n, v a n
[on]	[ɔ̃]	ombre, bon	o n b r, b o n
[un]	[œ̃]	lundi, brun	l u n d i, b r u n
[y]	[j]	yeux, pied	y e u, p y e
[w]	[w]	oui, nouer	w i, n w e
[yw]	[ɥ]	huile, lui	y w i l, l y w i
[p]	[p]	père	p a i r
[t]	[t]	terre	t a i r
[k]	[k]	cou, sac	k o u, s a k
[b]	[b]	bon, robe	b o n, r o b
[d]	[d]	dans, aide	d a n, a i d
[g]	[g]	gare, bague	g a r, b a g
[f]	[f]	feu, phare	f e u, f a r
[s]	[s]	sale, tasse	s a l, t a s
[sh]	[ʃ]	chat, tache	s h a, t a s h
[v]	[v]	vous, rêve	v o u, r a i v
[z]	[z]	zéro, maison	z e r o, m a i z o n
[j]	[ʒ]	je, gilet	j @, j i l a i
[l]	[l]	lent, sol	l a n, s o l
[r]	[ʁ]	rue, venir	r u, v @ n i r
[m]	[m]	main, femme	m i n, f a m
[n]	[n]	nous, tonne	n o u, t o n
[ny]	[ɲ]	agneau, vigne	a n y o o, v i n y
[h]	[h]	hop	h o p
[ng]	[ŋ]	camping	c a n p i n g
[x]	[x]	jotta	x o t a

TAB. G.1 – *Alphabet phonétique utilisé dans le lexique ENST*

dans une base contenant environ un million de noms propres. Les prononciations de cette base de données ont été produites par des méthodes semi-automatique, en particulier, elles n'ont pas toutes été vérifiées par un expert. Ces deux raisons font que, des quatre lexiques d'ONOMASTICA, ce lexique est probablement celui qui est le moins homogène, et qui présente, en conséquence, les plus grandes difficultés de transcription.

Les transcriptions de ce lexique sont décrites par un alphabet de 51 phonèmes.

Annexe H

Le lexique CMU

Le lexique CMU est un lexique anglais phonétisé du domaine public. Ce lexique a été développé à l'université de Carnegie Mellon, où il est utilisé dans un système de compréhension de parole. Nous avons utilisé la version 0.3 de ce lexique. Par rapport à NETTALK, ce lexique présente l'avantage d'être beaucoup plus volumineux (107 029 entrées), ce qui nous a permis de construire des courbes d'apprentissage asymptotiques sur des plages de valeurs beaucoup plus grandes. Ce lexique présente par contre l'inconvénient d'être moins « propre » que le lexique NETTALK, dans la mesure où, d'une part, il contient des transcriptions d'origine diverse, dont certaines n'ont pas été contrôlées manuellement; et d'autre part la notion d'entrée lexicale valide est un peu plus lâche que dans le lexique NETTALK. En particulier, dans sa version originale, il contient des nombres, des sigles, des formes cliticisées, etc, ainsi que des formes fléchies (pluriels, formes conjuguées).

Nous avons fait subir au lexique d'origine un certain nombre de pré-traitements, de manière notamment à normaliser les entrées. Ces traitements ont principalement consisté à :

- éliminer les entrées non-alphabétiques;
- développer les abréviations les plus fréquemment utilisées;
- éliminer les formes cliticisées;
- produire un alignement entre formes graphiques et phonétiques;

Une version abrégée de ce lexique, contenant un sous-ensemble aléatoirement sélectionné de 20 000 entrées, a également été utilisé dans nos expériences, sous le nom de CMU-20.

Le système de transcription phonétique utilisé dans le lexique CMU est assez différent de celui utilisé dans NETTALK. En particulier (voir table H.1), le niveau de description des voyelles est moins précis, puisqu'il ne fait appel qu'à 15 symboles pour les voyelles et diphtongues (contre 22 dans NETTALK). De manière à rendre la tâche d'apprentissage plus complexe, nous avons conservé dans les

transcriptions phonémiques, le niveau d'accent lexical associé à chaque voyelle (0, 1, ou 2). L'entrée lexicale correspondant au mot *tradition* dans notre base de données aura donc l'allure suivante ;:

tradition t r ah0 d ih1 sh ah0 n

Par ce biais, le nombre de symboles de l'alphabet phonétique du lexique CMU est en fait de 75, ce qui rend l'apprentissage de la transcription à partir de ces données plus difficile (en théorie, car la connaissance des positions accentuées est aussi une aide pour le système d'apprentissage) qu'à partir des données de NETTALK.

CMU	IPA	Comme dans ...	Transcrit ...
[iy]	[i:]	bead, see	b iy d, s iy
[aa]	[ɑ:]	bard, calm	b aa r d, k aa m
[ao]	[ɔ:]	born, cork	b ao r n, k ao r k
[uw]	[u:]	boon, fool	b uw n, f uw l
[er]	[ɜ:]	burn, work	b er n, w er k
[ih]	[ɪ]	sit, pity	s ih t, p ih t ih
[eh]	[e]	set, less	s eh t, l eh s
[ae]	[æ]	sat	s ae t, ae p ah l
[ah]	[ʌ]	fun, come	f ah n, k ah m
[uh]	[ʊ]	full, soot	f uh l, s u t
[ey]	[eɪ]	bay, fate	b ey, f ey t
[ay]	[aɪ]	buy, lie	b ay, l ay
[oy]	[ɔɪ]	voice, boy	v oy s, b oy
[ow]	[əʊ]	no, ago	n ow, ah g ow
[aw]	[aʊ]	now, plough	n aw, p l aw
[y]	[j]	yet, million	y eh t, m ih l y ah n
[w]	[w]	wet, bewail	w eh t, b i w ey l
[hh]	[h]	hat, preheat	hh ae t, p r iy hh iy t
[b]	[p]	pat, pope	p ae t, p ow p
[t]	[t]	tab, strut	t ae b, s t r ah t
[k]	[k]	cot, kiss	k aa t, k ih s
[b]	[b]	bat, baby	b ae t, b ey b iy
[d]	[d]	dab, amend	d ae b, ah m eh n d
[g]	[g]	got, agog	g aa t, ah g aa g
[f]	[f]	fine, raffle	f ay n, r ae f ah l
[s]	[s]	sit, rice	s ih t, r ay s
[sh]	[ʃ]	ship, sugar	sh ih p, sh uh g er
[th]	[θ]	thin, math	th ih n, m ae th
[ch]	[tʃ]	chance	ch ae n s
[v]	[v]	vine, river	v ay n, r ih v er
[z]	[z]	pod, buzz	p aa d z, b ah z
[zh]	[ʒ]	measure	m eh zh er
[dh]	[ð]	this, other	dh ih s, ah dh er
[jh]	[dʒ]	just, edge	jh ah s t, eh jh
[l]	[l]	little, place	l ih t ah l, p l ey s
[r]	[r]	ran, steer	r ah n, s t ih r
[m]	[m]	ram, mammoth	r ae m, m ae m ah th
[n]	[n]	ran, nut	r ah n, n ah t
[ng]	[ŋ]	rang, bank	r ae ng, b ae n ng

TAB. H.1 – *Alphabet phonétique utilisé dans le lexique CMU (les accents ont été omis)*

Annexe I

Le lexique de verbes anglais

Le lexique de verbes anglais utilisé pour les expériences concernant la prédiction du passé a été construit par MacWhinney et Leinbach à partir du corpus de Francis et Kucera. De ce corpus, les 6939 formes verbales (présent (3^e personne et autres), gérondif, passé, participe passé) les plus fréquentes ont été extraites, correspondant à 2161 verbes différents. Ce lexique est accessible à l'adresse ftp suivante : poppy.psy.cmu.edu, les données étant dans le répertoire /pdp/.

La répartition par flexion est donnée dans la table I.1.

Forme	Fréquence
Présent (hors 3 ^e pers. sing.)	2163
Présent (3 ^e pers. sing)	632
Participe présent	1325
Passé	1394
Participe passé	1425
Total	6939

TAB. I.1 – Répartition des formes verbales par flexion

Ce lexique contient, pour chaque forme verbale, les informations suivantes :

- la graphie de la forme radicale du verbe;
- la graphie de la forme fléchie;
- le type de flexion associé à l'entrée courante;
- Une transcription phonétique dans l'alphabet Unibet (voir la table I.2) de la forme radicale;
- une transcription phonétique dans l'alphabet Unibet de la forme fléchie;
- la fréquence de la forme verbale dans le corpus de Francis et Kucera;

- un indicateur (régulier/irrégulier) de la régularité de la forme;

Nous avons utilisé ce lexique tel que, sans notamment supprimer les formes multiples d'un même verbe lorsque deux conjugaisons sont acceptables, ni les formes homographes ou homophones de verbes différents.

La première série d'expériences (voir la section 4.5.2) les couples d'exemples en entrée sont de la forme:(transcription phonétique du radical, transcription phonétique de la forme fléchie), comme par exemple: (b&t6r,b&t6rd), correspondant au verbe *batter*.

Pour la seconde série d'expérience (section 4.5.2), nous avons eu recours à une représentation abstraite des formes fléchies. L'exemple correspondant au verbe *batter* correspond alors à deux entrées dans le fichier d'exemples: (b&t6r+B,b&t6r) et (b&t6r+D,b&t6rd).

L'expérience d'analyse (décrite également à la section 4.5.2) utilise les mêmes données que la précédente, en inversant simplement les entrées et les sorties.

Unibet	API	As in ...	Transcribed as ...
[i]	[i]	bead, see	bid, si
[a]	[ɑ]	bard, calm	bard, kam
[o]	[ɔ]	born, cork	born, kork
[u]	[u]	boon, fool	bun, ful
[6r]	[ɜ]	burn, work	b6rn, w6rk
[I]	[ɪ]	sit, pity	sIt, pItY
[E]	[e]	set, less	sEt, lEs
[&]	[æ]	sat, apple	s&t, &pl
[6]	[ʌ]	fun, come	f^n, k^m
[O]	[ɒ]	fond, wash	fOnd, wOS
[U]	[ʊ]	full, soot	fUl, sUt
[6]	[ə]	composer, above	k6mpoz6r, 6b6v
[e]	[et]	bay, fate	be, fet
[3]	[aɪ]	buy, lie	b3, l3
[2]	[ɔɪ]	voice, boy	v2s, b2
[o]	[aʊ]	now, plough	no, plo
[I]	[Iə]	tier, bier	tIr, bIr
[&]	[eə]	tare, fair	t&r, f&r
[j]	[j]	yet, million	jEt, mIlj6n
[w]	[w]	wet, bewail	wEt, bIwel
[h]	[h]	hat, preheat	h&t, prihit
[p]	[p]	pat, pope	p&t, pop
[t]	[t]	tab, strut	t&b, str6t
[k]	[k]	cot, kiss	kat, kIs
[b]	[b]	bat, baby	b&t, bebi
[d]	[d]	dab, amend	d&b, 6mEnd
[g]	[g]	got, agog	gat, 6gag
[f]	[f]	fine, raffle	f3n, r&fL
[s]	[s]	sit, rice	sIt, r3s
[S]	[ʃ]	ship, sugar	SIp, SUG6r
[T]	[θ]	thin, math	TIn, m&T
[C]	[tʃ]	chance	C&ns
[v]	[v]	vine, river	v3n, rIv6r
[z]	[z]	Pods, buzz	padz, b6z
[Z]	[ʒ]	measure	mEZ6r
[D]	[ð]	this, other	DIs, 6D6r
[l]	[l]	little, place	lItl, ples
[r]	[r]	ran, steer	r6n, stIr
[m]	[m]	ram, mammoth	r&m, m&m6T
[n]	[n]	ran, nut	r6n, n6t
[N]	[ŋ]	rang, bank	r&N, b&Nk

TAB. I.2 – *Alphabet phonétique Unibet utilisé dans le corpus de verbes anglais*

Références

- ADRIAANS, P., ed. 1993. *Proceedings of the ICML workshop on Machine Learning and Text Analysis*.
- AHA, D.; KIBLER, D. et ALBERT, M. 1991. Instance-based learning algorithm. *Machine Learning* 6:37-66.
- AINSWORTH, W. A. et PELL, B. 1989. Connectionist architectures for a text-to-speech system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, 125-128.
- ANDERSEN, O. et DALSGAARD, P. 1993. A self-learning approach to transcription of proper names. Draft research paper from the LRE project ONOMASTICA.
- AUBERGÉ, V. 1991. *La synthèse de la parole: « des règles au lexique »*. Ph.D. Dissertation, Université Stendhal, Grenoble.
- BAKIRI, G. et DIETTERICH, T. G. 1993. Performance comparison between human engineered and machine learned letter-to-sound rules for English: A machine learning success story. In *Proceedings of the 18th International Conference on the applications of Computer and Statistics to Science and Society*.
- BARTKOVA, K.; LARREUR, D. et MÉTAYER, I. 1994. Choix et adaptation d'un phonétiseur pour la reconnaissance automatique de la parole. In *Actes des XXème Journées d'Études sur la Parole*, 181-186.
- BAUM, L. E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process. *Inequalities* 3, 1-8.
- BEAUDOIN, V. et YVON, F. 1994. Le métromètre, ou le décompte des positions par lesquelles le vers se fait mètre. *Mezura* (32).
- BEAUDOIN, V. et YVON, F. 1996. The metrometer: a tool for analysing the French verse. *Literary and Linguistic Computing* 11(1).
- BÉCHET, F.; DERDERIAN, S. et EL-BÈZE, M. 1995. Conversion graphèmes-phonèmes automatique: le système Grifon. In *Actes des 15èmes journées internationales IA'95*.
- BELHOULA, K. 1993. Rule-based grapheme-to-phoneme conversion of names. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 881-884.
- BESLING, S. 1995. A statistical approach to multilingual phonetic transcription. *Philips Journal of Research*.
- BIMBOT, F.; PIERRACCINI, R.; EVIN, E. et ATAL, B. 1994. Modèles de séquence à horizon variable: multigrammes. In *Actes des XXèmes journées d'études sur la parole*, 467-472.
- BIMBOT, F.; DELIGNE, S. et YVON, F. 1995. Unsupervised decomposition of phoneme strings into variable-length sequences, by multigrams. In *Proceedings of the International Congress of Phonetic Science (ICPhS)*.

- BOD, R. 1995. *Enriching Linguistics with Statistics: Performance Models of Natural Language*. Institute for Logic, Language and Computation, Amsterdam.
- BOULA DE MAREUIL, P. 1994. Vers une phonémisation automatique des sigles. In *Actes des XXème Journées d'Études sur la Parole*, 95-100.
- BROWN, P. et BESNER, D. 1986. The assembly of phonology in oral reading: a new model. In COLTHEART, M., ed., *Attention and performance XII: The psychology of reading*. London, UK: Lawrence Erlbaum Associates.
- BULLINARIA, J. A. 1995. Neural network learning from ambiguous training data. *Connection Science* 7(2):99-122.
- BYBEE, J. L. 1988. Morphology as lexical organization. In *Theoretical Morphology*. Academic Press. chapter 7, 119-141.
- BYRD, R. J. et CHODOROW, M. S. 1985. Using an on-line dictionary to find rhyming words and pronunciations for unknown words. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 277-283.
- CALIFF, M. et MOONEY, R. J. 1995. Using inductive logic programming to learn the past tense of English verbs. In WERMTER, S., ed., *Proceeding of the IJCAI'95 workshop on 'New Approaches to Learning for Natural Language Processing'*, 87-94.
- CARAMAZZA, A.; LAUDANNA, A. et ROMANI, C. 1988. Lexical access and inflectional morphology. *Cognition* 28:297-332.
- CARLSON, R.; GRANSTRÖM, B. et LINDSTRÖM, A. 1989. Predicting name pronunciation for a reverse directory service. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 113-116.
- CARTWRIGHT, T. A. et BRENT, M. R. 94. Segmenting speech without a lexicon: the roles of phonotactics and speech source. In *Proceedings of the 1st meeting of the Association for Computational Phonology*.
- CASALIS, S. et LECOCQ, P. 1992. Les dyslexies. In *Psychologie cognitive de la lecture*. Paris: Presses Universitaires de France. 195-237.
- CATACH, N. 1984. *La phonétisation automatique du Français*. Paris: Éditions du CNRS.
- CHARNIAK, E. 1993. *Statistical Language Learning*. Cambridge, MA: The MIT Press.
- CHARON, I. et HUDRY, O. 1993. The noising method: a new method for combinatorial optimization. *Optimization research letter* 14:133-137.
- CHOMSKY, N. et HALLE, M. 1968. *The Sound Pattern of English*. Cambridge, Massachussets: The MIT Press.
- CHURCH, K. 1985. Stress assignment in letter to sound rules for speech synthesis. In *Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 246-253.
- COKER, C. H.; CHURCH, K. W. et LIBERMAN, M. Y. 1990. Morphology and rhyming: two powerful alternatives to letter-to-sound rules. In *Proceedings of the ESCA Conference on Speech Synthesis*.

- COLTHEART, M.; CURTIS, B.; ATKINS, P. et HALLER, M. 1993. Models of reading aloud: dual route and parallel distributed processing approaches. *Psychological Review* 100:589-608.
- COLTHEART, M. 1978. Lexical access in simple reading tasks. In UNDERWOOD, G., ed., *Strategies of information processing*. New York: Academic Press. 151-216.
- CONSTANT, P. 1991. *Analyse syntaxique par couche*. Ph.D. Dissertation, École Nationale Supérieure des Télécommunications (ENST), Paris, France.
- DAELEMANS, W. et POWERS, D., eds. 1992. *Proceedings of SHOE workshop: Background and Experiments in Machine Learning of Natural Language*.
- DAELEMANS, W.; GILLIS, S.; DURIEUX, G. et VAN DEN BOSCH, A. 1993. Learnability and markedness in data-driven acquisition of stress. Technical Report 43, ITK, Tilburg University, NL.
- DAELEMANS, W.; GILLIS, S. et DURIEUX, G. 1994a. The acquisition of stress: a data-oriented approach. *Computational Linguistics* 20(3):421-451.
- DAELEMANS, W.; GILLIS, S. et DURIEUX, G. 1994b. Skousen's analogical modeling algorithm: A comparison with lazy learning. In *Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP I)*.
- DAGAN, I. et ENGELSON, S. P. 1995. Selective sampling in natural language learning. In *Proceedings of IJCAI'95 workshop on 'New Approaches to Learning for Natural Language Processing'*.
- DAUZAT, A. et ROSTAING, C. 1963. *Dictionnaire étymologique des noms de lieux en France*. Paris: Guénégaud, 1978, 2ème édition.
- DAUZAT, A. 1949. *Les noms de famille en France*. Paris: Guénégaud, 1988, 3ème édition.
- DE SAUSSURE, F. 1916. *Cours de Linguistique Générale*. Paris: Payot.
- DEDINA, M. J. et NUSBAUM, H. C. 1991. PRONOUNCE: a program for pronunciation by analogy. *Computer Speech and Language* 5:55-64.
- DELIGNE, S. et BIMBOT, F. 1995. Language modelling by variable length sequences: theoretical formulation and evaluation of multigrams. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- DELIGNE, S.; YVON, F. et BIMBOT, F. 1995. Phonetic transcription by variable length sequences: joint multigrams. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.
- DELIGNE, S. 1996. *À paraître*. Ph.D. Dissertation, Thèse de l'École Nationale Supérieure des Télécommunications.
- DELL, F. 1985. *Les règles et les sons*. Paris: Hermann, 2ème édition.
- DELL, F. 1995. Consonant clusters and phonological syllables in French. *Lingua* 95.
- DEMPSTER, A. P.; LAIRD, N. M. et RUBIN, D. B. 1977. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* 39(1):1-38.

- DIAMOND, K. 1994. L'oralisation des sigles en anglais. *Lynx* (30):109-132.
- DIETTERICH, T. G. et BAKIRI, G. 1991. Error-correcting output codes: a general method for improving multiclass inductive learning programs. In *Proceedings of the meeting of the American Association for Artificial Intelligence (AAAI)*, 572-577.
- DIETTERICH, T. G.; HILD, H. et BAKIRI, G. 1990. A comparative study of ID3 and back-propagation for English text-to-speech mapping. In *Proceedings of the 7th International Machine Learning Workshop*. Austin: Morgan Kaufman.
- DIETTERICH, T. G.; HILD, H. et BAKIRI, G. 1995. A comparison of ID3 and backpropagation for English text-to-speech mapping. *Machine Learning* 18(1):51-80.
- DIVAY, M. 1990. A written text processing expert system for text to phoneme conversion. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, 853-856.
- DURAND, J. 1990. *Generative and Non-linear Phonology*. London: Longman.
- DURIEUX, G. 1992. Analogical modelling of main stress assignment in Dutch simplex words. In DAELEMANS, W. et POWERS, D., eds., *Proceedings of the first SHOE workshop*, 197-203.
- ENCREVÉ, P. 1988. *La Liaison avec et sans Enchaînement*. Paris: Editions du Seuil.
- FALKENHAINER, B.; FORBUS, K. D. et GENTNER, D. 1990. Structure mapping engine. *Artificial Intelligence* 41.
- FAYOL, M. 1992. Comprendre ce que l'on lit: de l'automatisme au contrôle. In *Psychologie cognitive de la lecture*. Paris: Presses Universitaires de France. 73-105.
- FEDERICI, S. et PIRRELLI, V. 1994a. Analogy as a computable process. In *Proceedings of the conference on New Methods in Natural Language Processing (NeMLap I)*. Manchester: Umist.
- FEDERICI, S. et PIRRELLI, V. 1994b. The compilation of large pronunciation lexica: the elicitation of letter-to-sound patterns through analogy-based networks. In *Proceedings of COMPLEX 94*.
- FEDERICI, S.; PIRRELLI, V. et YVON, F. 1995. Advances in analogy-based learning: false friends and exceptional items in pronunciation by paradigm-driven analogy. In *Proceedings of IJCAI'95 workshop on 'New Approaches to Learning for Natural Language Processing'*.
- FEDERICI, S.; PIRRELLI, V. et YVON, F. 1996. A dynamic approach to paradigm-driven analogy. In *Symbolic, connectionist, and statistical approaches to learning for natural language processing*. Berlin: Springer-Verlag.
- FORBUS, K. D.; GENTNER, D. et LAW, K. 1994. MAC/FAC: A model of similarity-based retrieval. *Cognitive Science* 19:141-205.
- FORSTER, K. I. et CHAMBERS, S. M. 1973. Lexical access and naming time. *Journal of verbal learning and verbal behavior* 12:627-635.

- FORSTER, K. I. 1990. Lexical processing. In OSHERSON, D. N. et LASNIK, H., eds., *An invitation to cognitive science: Language*, 94-131. Cambridge, MA: The MIT Press.
- FOUCHÉ, P. 1959. *Traité de prononciation française*. Paris: Klincksieck.
- FRANÇOIS, D. 1968. Les argots. In *Le Langage*, 620-646. Paris: Gallimard.
- GAVED, M. 1993. Pronunciation and text normalization in applied text-to-speech systems. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, 897-900.
- GILLIS, S.; DAELEMANS, W. et DURIEUX, G. 1994. Are children 'lazy learner'? A comparison of natural and machine learning of stress. In *Proceedings of the sixteenth Annual Conference of the Cognitive Science Society*, 369-374.
- GLUSHKO, J. R. 1979. The organization and activation of orthographic knowledge in reading aloud. *Journal of experimental psychology: Human perception and performance* 5:674-691.
- GLUSHKO, J. R. 1981. Principles for pronouncing print: the psychology of phonography. In LESGOLD, A. M. et PERFETTI, C. A., eds., *Interactive Processes in Reading*, 61-84. Hillsdale, New Jersey: Erlbaum.
- GOLDING, A. R. et ROSENBLOOM, P. S. 1991. Improving rule-based systems through case-based reasoning. In *Proceedings of the meeting of the American Association for Artificial Intelligence (AAAI)*, 22-27.
- GOLDING, A. R. 1991. *Pronouncing names by a combination of rule-based and case-based reasoning*. Ph.D. Dissertation, Stanford University, Stanford, CA.
- GROSS, M. et LENTIN, A. 1970. *Notions sur les grammaires formelles*. Paris: Gauthier-Villars.
- GRUAZ, C. 1995. Les variantes graphiques du français: système et arbitraire dans la synchronie contemporaine. *Langue Française* (108):18-24.
- GRUNE, D. et JACOB, C. J. 1990. *Parsing Techniques: a practical Guide*. Ellis Horwood.
- HATON, J.-P.; BOUZID, N.; FRANÇOISCHARPILLET; HATON, M.-C.; LÂASRI, B.; LÂASRI, H.; MARQUIS, P.; MONDOT, T. et NAPOLI, A. 1991. *Le raisonnement en Intelligence Artificielle*. Interéditions.
- HENSEN, A. V. 94. A self-organizing neural network (SONN). In *Proceedings of the second Onomastica Research Colloquium*.
- HOCHBERG, J.; MNISZEWSKI, S.; CALLEJA, T. et PAPCUN, G. J. 1990. What's in a name? Last names as a computational problem. Unpublished paper, Los Alamos National Laboratory, Los Alamos, NM.
- HOFFMAN, R. R. 1995. Monster analogies. *AI Magazine* 11-35.
- HOPCROFT, J. E. et ULLMAN, J. D. 1979. *Introduction to automata theory, languages and computation*. Addison-Wesley.
- HUNNICUTT, S.; MENG, H.; SENEFF, S. et ZUE, V. W. 1993. Reversible letter-to-sound sound-to-letter generation based on parsing word morphology. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, 763-766.

- INDURKHYA, B. 1992. *Metaphor and Cognition: an interactionist approach*. Kluwer Academic Publishers.
- JELINEK, F. et MERCER, M. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, 381-397.
- JONES, D., ed. 1994. *Proceedings of the conference on New Methods in Natural Language Processing (NeMLaP I)*.
- KOHONEN, T. 1984. *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.
- KOHONEN, T. 1986. Dynamically expanding context with application to the correction of symbol strings in the recognition of continuous speech. In *Proceedings of the 8th International Conference on Pattern Recognition*, volume 2, 1148-1151.
- KOLODNER, J. 1993. *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.
- LAGHERET-DUJOUR, A. 1990. *Contribution à une analyse de la variabilité phonologique pour le traitement automatique de la parole continue multi-locuteur*. Ph.D. Dissertation, Université de Paris VII.
- LACHTER, J. et BEVER, T. 1988. The relation between linguistic structure and associative theories of language learning: a constructive critic of some connectionist learning models. *Cognition* 28:195-247.
- LAKOFF, G. et JOHNSON, H. 1985. *Les métaphores dans la vie quotidienne*. Éditions de Minuit.
- LAPORTE, É. 1988a. *Méthodes Algorithmiques et Lexicales de Phonétisation de Textes*. Ph.D. Dissertation, Université de PARIS VII.
- LAPORTE, É. 1988b. Phonétisation de textes par un réseau connexionniste. Technical report, LIMSI.
- LAPORTE, É. 1989. Quelques variations phonétiques en français. *Linguisticae Investigationes* XII(1):43-116.
- LAPORTE, É. 1992. Phonetic syllables in French : combinatorics, structure and formal definitions. *Acta Linguistica Hungarica* 41:175-189.
- LAWRENCE, S. et KAYE, G. 1986. Alignment of phonemes with their corresponding orthography. In *Proceedings of the Autumn Conference of the Institute of Acoustics*, volume 8, 461-469.
- LEHNERT, W. G. 1987. Case-based problem solving with a large knowledge base of learned cases. In *Proceedings of the meeting of the American Association for Artificial Intelligence (AAAI)*, 301-306.
- LEMEUR, P.-Y. 1996. *Synthèse de la parole par recombinaison d'unités de taille variable*. Ph.D. Dissertation, École Nationale Supérieure des Télécommunications, Paris.
- LING, C. X. et MARINOV, M. 1993. Answering the connectionist challenge: a symbolic model of learning the past tense of english. *Cognition* 49(3):235-290.

- LING, C. X. 1994. Learning the past tense of english: the symbolic pattern associator vs. connectionist models. *Journal of Artificial Intelligence Research* 1:209-229.
- LLORACH, E. 1968. Communication orale et graphique. In *Le Langage*, 515-567. Paris: Gallimard.
- LUCAS, S. M. et DAMPER, R. I. 1992. Syntactic neural networks for bi-directional text-phonetics translation. In BAILLY, G. et BENOIT, C., eds., *Talking Machines*, 127-141. North Holland.
- LUCASSEN, J. et MERCER, R. 1984. An information theoretic approach to the automatic determination of phonemic baseforms. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, 45.5.1-42.5.4.
- LUK, R. W. et DAMPER, R. I. 1991. A novel approach to inferring letter-phoneme correspondences. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 741-744.
- LUK, R. W. et DAMPER, R. I. 1992. Inference of letter-phoneme correspondences by delimiting and dynamic time warping techniques. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 61-64.
- LUK, R. W. et DAMPER, R. I. 1993. Inference of letter-phoneme correspondences with pre-defined consonant and vowel patterns. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 203-206.
- MACWHINNEY, B. et LEINBACH, J. 1991. Implementations are not conceptualizations: revising the verb learning model. *Cognition* 40:121-158.
- MACWHINNEY, B. 1993. Connections and symbols: closing the gap. *Cognition* 49(3):291-296.
- MAES, P. 1993. *La Prononciations des langues européennes*. Paris: Éditions du Centre de Formation Professionnel des Journalistes.
- MARCUS, S. 1967. *Introduction mathématique à la linguistique structurale*. Paris: Dunod.
- MARTINET, A. 1967. *Éléments de linguistique générale*. Paris: Armand Colin, 3ème édition.
- MCCULLOCH, N.; BEDWORTH, M. et BRIDLE, J. 1987. NETspeak - a reimplementation of NETtalk. *Computer Speech and Langage* 2:289-302.
- MCCULLY, C. et HOLMES, M. 1988. Some notes on the structure of acronyms. *Lingua* 74(1):27-43.
- MENG, H.; SENEFF, S. et ZUE, V. W. 1994a. Phonological parsing for reversible letter-to-sound/sound-to-letter generation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, II-1-II-4.
- MENG, H.; SENEFF, S. et ZUE, V. W. 1994b. Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation. In *Proceedings ARPA Human Langage Technologies Workshop*.

- MENG, H.; SENEFF, S. et ZUE, V. W. 1994c. The use of higher level linguistic knowledge for letter-to-sound generation. In *Proceedings of the International Symposium on Speech, Image Processing and Neural Networks*, 670-673.
- MÉTAYER, I. et VINCENT, M. 91. La synthèse vocale des noms propres : problèmes et propositions de solutions. Technical report, CNET.
- MICLET, L. 1979. *Inférence de grammaires régulières*. Ph.D. Dissertation, École Nationale Supérieure des Télécommunications, Paris.
- MITCHELL, M. et HOFSTADTER, D. R. 1990. The emergence of understanding in a computer model of concepts and analogy-making. *Physica* 42:322-334.
- MOLINIER, C. 1992. Sur la productivité adverbiale des adjectifs. *Langue Française* (96):65-73.
- MOONEY, R.; SHAVLIK, J.; TOWELL, G. et GOVE, A. 1989. An experimental comparison of symbolic and connectionist algorithms. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI)*, volume 1, 775-780.
- MORTUREUX, M.-F. 1994. Siglaison-acronymie et néologie lexicale. *Lynx* (30):11-32.
- MOUNIN, G. 1974. *Dictionnaire de la linguistique*. Paris: Presses Universitaires de France.
- Oakey, S. et Cawthorn, R. 1981. Inductive learning of pronunciation rules by hypothesis testing and correction. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 109-114.
- PARFITT, S. et SHARMAN, R. 1991. A bi-directional model of English pronunciation. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, 801-804.
- PECATTE, J.-M. 1992. *Tolérance aux fautes dans les interfaces homme-machine. Traitement des chaînes phonétiques, des chaînes orthographiques et des structures syntaxiques*. Ph.D. Dissertation, Université Paul Sabatier de Toulouse III.
- PÉRENNOU, G.; COTTO, D.; DE CALMÈS, M.; FERRANÉ, I.; PÉCATTE, J.-M. et TIHONI, J. 1991. Composantes phonologique et orthographique de BDLEX. In *Actes des 2èmes journées du GRECO-PRC CHM*. Toulouse, France: EC2 Editeur.
- PINKER, S. et PRINCE, A. 1988. On language and connectionism: analysis of a parallel distribution processing model of language acquisition. *Cognition* 28:73-194.
- PINKER, S. 1994. *The language instinct*. Penguin Books.
- PIRRELLI, V. et FEDERICI, S. 1992. An analogical way to language modelling: Morpheus. *Acta Linguistica Hungarica* 41:235-263.
- PIRRELLI, V. et FEDERICI, S. 1994a. Context-sensitivity and linguistic structure in analogy-based parallel networks. In *Current issues in Mathematical Linguistics*, 343-362. Elsevier.

- PIRRELLI, V. et FEDERICI, S. 1994b. "Derivational" paradigms in morphology. In *Proceedings of the sixteenth International Conference on Computational Linguistics (COLING)*.
- PIRRELLI, V. et FEDERICI, S. 1994c. On the pronunciation of unknown words by analogy in text-to-speech systems: an evaluation. In *Proceedings of the 2nd Onomastica Research Colloquium*.
- PIRRELLI, V. et FEDERICI, S. 1995. "You'd better nothing than say something wrong": analogy, accuracy and text-to-speech applications. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.
- PIRRELLI, V. et SALZA, P. 1993. Transcription rules and stress assignment in ONOMASTICA. Technical Report R-001, ILC/CSELT.
- PLAUT, D. et MCLELLAND, J. L. 1993. Generalization with componential attractors: word and non-word reading in an attractor network. In *Proceedings of the 15th annual meeting of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- PLÉNAT, M. 1991. Vers une oralisation des sigles. In *Actes des IIeme Journées internationales du GRECO-PRC Communication Homme-Machine*, 363-371. Toulouse: EC2 Eds, Nanterre.
- PLÉNAT, M. 1992. Quelques problèmes touchant l'oralisation des sigles. Présentation au Séminaire Lexique IRT-UPS, Toulouse.
- PLÉNAT, M. 1993. Remarques sur la structure de la rime en français standard. In PERRET, A. E. . M., ed., *Etudes de Linguistique Française à la Mémoire d'Alain Lerond*. La Garenne-Colombe: Editions de l'Espace Européen. 261-288.
- PLÉNAT, M. 1994. Observations sur le mot minimal français. In LAKS, B. et PLÉNAT, M., eds., *De Natura Sonorum*. Saint-Denis: Presses Universitaires de Vincennes. 144-172.
- PLÉNAT, M. 1996. De l'interaction des contraintes: une étude de cas. In DURAND, J. et LAKS, B., eds., *Current trends in phonology*. CNRS Paris X and University of Salford.
- PLUNKETT, K. et MARCHMAN, V. 1990. From rote learning to system building: acquiring verb morphology in children and connectionist nets. Technical Report 9020, Center for Research in Language, University of California, San Diego, CA.
- POWERS, D. et REEKER, L., eds. 1991. *Proceedings of the workshop on Machine Learning of Natural Languages and Ontology (MLNLO)*.
- PRINCE, A. et SMOLENSKY, P. 1992. Optimality theory: constraint interaction in generative grammar. Technical Report 2, the Rutgers Center for Cognitive Science.
- QUINLAN, J. R. 1983. Learning efficient classification procedures and their applications to chess end games. In MICHALSKI, R. S.; CARBONELL, J. G. et MITCHELL, T. M., eds., *Machine Learning I*. Palo Alto, CA: Tioga Publishing Company. chapter 15, 463-482.
- QUINLAN, J. R. 1987. Simplifying decision trees. *International Journal of Man-Machine Studies* 27:221-234.

- QUINLAN, J. R. 1993. *C4.5 Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.
- RABINER, L. et JUANG, B. 1993. *Fundamentals of Speech Recognition*. Englewoods Cliff, NJ: PTR Prentice Hall.
- RAJMAN, M. 1995. *Apports d'une approche à base de corpus aux techniques de traitement automatique du langage naturel*. Ph.D. Dissertation, École Nationale Supérieure des Télécommunications, Paris.
- RENTZEPOPOULOS, P. A. et KOKKINAKIS, G. K. 1991. Phoneme to grapheme conversion using HMM. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, 797-800.
- ROSSON, M. 1985. The interaction of pronunciation rules and lexical representations in reading aloud. *Memory and Cognition* 13:90-99.
- RUMELHART, D. E. et MCCLELLAND, J. L. 1986. On learning the past tenses of english verbs. In MCCLELLAND, J. L. et RUMELHART, D. E., eds., *Parallel distributed processes: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models*. Cambridge, Massachussets: Bradford books/MIT Press. 217-271.
- SAGISAKA, Y. et IWAHASHI, N. 1995. Objective optimization in algorithms for text-to-speech synthesis. In W.B, K. et PALIWAL, K., eds., *Speech Coding and Synthesis*. Elsevier Science B.V. 686-706.
- SCHMIDT, M.; FITT, S.; SCOTT, T. et JACK, M. 1993. Phonetic transcription standards for european names (ONOMASTICA). In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, 279-282.
- SEIDENBERG, M. S. et MCCLELLAND, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological review* 96:523-568.
- SEJNOWSKI, T. J. et ROSENBERG, C. R. 1987. Parrallel network that learn to pronounce English text. *Complex Systems* 1:145-168.
- SKOUSEN, R. 1989. *Analogical Modelling of Language*. Dordrecht: Kluwer.
- SORIN, C. 1991. Synthèse de la parole à partir du texte : état des recherches et des applications. In *Actes des IIeme Journées internationales du GRECO-PRC Communication Homme-Machine*. Toulouse: EC2 Eds, Nanterre.
- SPENCER, A. 1988. Bracketing paradoxes and the english lexicon. *Language* 64:663-682.
- SPENCER, A. 1991. *Morphological Theory*. Oxford, UK: Blackwell.
- SPIEGEL, M. F. et MACHI, M. J. 1990. Synthesis of names by a demisyllable-based speech synthesizer (ORATOR). *Journal of the American Voice Input/Output Society* 7:1-10.
- SPRENGER-CHAROLLES, L. 1992. L'évolution des mécanismes d'identification des mots. In *Psychologie cognitive de la lecture*. Paris: Presses Universitaires de France. 141-193.
- STANFILL, C. W. et WALTZ, D. 1986. Toward memory-based reasoning. *Communications of the ACM* 29:1213-1229.

- STANFILL, C. W. 1987. Memory-based reasoning applied to English pronunciation. In *Proceedings of the meeting of the American Association for Artificial Intelligence (AAAI)*, 577-581.
- SULLIVAN, K. et DAMPER, R. I. 1990. Psychologically-governed approach to novel-word pronunciation within a text-to-speech system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 341-344.
- SULLIVAN, K. et DAMPER, R. I. 1991. Speech synthesis by analogy: recent advances and results. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 761-764.
- SULLIVAN, K. et DAMPER, R. I. 1992. Synthesis-by-analogy: a bi-lingual investigation using German and English. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, 113-116.
- TARABAN, R. et MCLELLAND, J. L. 1987. Conspiracy effects in word pronunciation. *Journal of Memory and Language* 26:608-631.
- THAGARD, P.; HOLOYAK, K. J.; NELSON, G. et GOCHFELD, D. 1990. Analog retrieval by constraint satisfaction. *Artificial Intelligence* 46(3):259-310.
- TIHONI, J. et PÉRENNOU, G. 1991. Phonotypical transcription through the GEPH expert system. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 767-770.
- TORKOLLA, K. 1993. An efficient way to learn English grapheme-to-phoneme rules automatically. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 199-202.
- TREIMAND, R. et ZUKOWSKI, A. 1988. Units in reading and spelling. *Journal of Memory and Language* 27:466-477.
- TZOUKERMAN, E. 1993. Progress in French text-to-speech systems. Technical report, AT&T Bell Laboratories.
- VAN COILE, B.; LEYS, S. et MORTIER, L. 1992. On the development of a name pronunciation system. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 1, 487-490.
- VAN COILE, B. 1991. Inductive learning of pronunciation rules with the DEPES system. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, 745-748.
- VAN COILE, B. 1993. On the development of pronunciation rules for text to speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 2, 1455-1458.
- VAN DEN BOSCH, A. et DAELEMANS, W. 1992. Generalization performance of backpropagation learning on a syllabification task. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*.
- VAN DEN BOSCH, A. et DAELEMANS, W. 1993. Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, 45-53.
- VIANA, M.; TRANCOSO, I. M. et SILVA, F. M. 1994. On the pronunciation of proper names and acronyms in european portugese. In *Proceedings of the second ONOMASTICA research colloquium*.

- VIDAL, E.; PIERACCINI, R. et LEVIN, E. 1993. Learning associations between grammars: a new approach to natural language understanding. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*.
- VITALE, T. 1991. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics* 17(3):257-276.
- VITERBI, A. 1967. Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory* 13:260-269.
- WARNANT, L. 1987. *Dictionnaire de la prononciation française*. Paris: Duculot.
- WERMTER, S. et RILOFF, E., eds. 1995. *Proceedings of the IJCAI'95 workshop on 'New Approaches to Learning for Natural Language Processing'*.
- WIOLAND, F. 1985. *Les Structures Syllabiques du Français*. Genève: Slatkine-Champion.
- WU, D. 1995. Stochastic inversion transduction grammars, with application to segmentation, bracketing, and alignment of parallel corpora. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 2, 1329-1335.
- YVON, F. 1993. A tidy grapheme to phoneme transcriber for ONOMASTICA. In *Proceedings of the Onomastica Research Colloquium*, 47-56.
- YVON, F. 1994a. Règles de transcription graphème-phonème pour la transcription automatique des sigles. *Lynx* (30):153-166.
- YVON, F. 1994b. Self-learning techniques for grapheme-to-phoneme conversion. In *Proceedings of the 2nd Onomastica Research Colloquium*.

Prononcer par analogie : motivation, formalisation et évaluation

La transcription d'un texte écrit en la séquence phonémique représentant (plus ou moins fidèlement) sa prononciation est une étape importante dans de nombreuses applications du traitement automatique du langage et de la parole. Cette opération est traditionnellement réalisée par des systèmes de règles de réécriture dépendant du contexte, auxquels sont adjoints, dans des proportions variables, des dictionnaires stockant les formes phonétiques des lexies les plus communes et/ou les plus irrégulières.

Tirant les conséquences de la grande difficulté de conception que posent ces systèmes de règles, des limites intrinsèques de l'approche par règle, et, par contraste, de la disponibilité croissante de vastes lexiques phonétiques, en particulier de noms propres, ainsi que de procédures permettant de les exploiter efficacement, ce travail envisage le développement de méthodes d'apprentissage de systèmes de transcriptions à partir de connaissances lexicales.

Nous procédons dans un premier temps à une revue critique de des multiples approches « classificatoires » de ce problème, qui considèrent cet apprentissage sous l'angle de l'inférence de fonctions discriminantes statistiques (réseaux de neurones, arbres de décision, plus proches voisins, etc). Dans une telle optique, le lexique est essentiellement considéré comme un ensemble de lettres, à partir duquel il s'agit d'inférer des règles de classement des symboles graphiques. Un prérequis indispensable à la mise en œuvre d'une tel apprentissage est que les représentations lexicales contiennent un appariement bi-univoque avec le phonème correspondant. Les problèmes pratiques et théoriques que ces approches soulèvent sont passés en revue et illustrés par des résultats obtenus à l'aide d'une méthode d'extraction d'arbres de décision.

Une première alternative aux systèmes classificatoires est ensuite développée, qui consiste à envisager la transcription des mots hors-lexique par recombinaison (exemplairement la juxtaposition) de fragments de taille variable d'entrées lexicales. Nous étendons les algorithmes existants, et proposons plusieurs mécanismes de construction et d'évaluation des multiples recombinaisons envisageables. Le principal résultat de cette étude est la définition d'un critère **structurel** d'évaluation des transcriptions, fondé sur le concept de « recouvrement » entre fragments adjacents, et à l'aide duquel nous obtenons des performances significativement meilleures que celles obtenues avec un algorithme classificatoire de référence.

Une seconde alternative aux systèmes classificatoires est finalement présentée, au fondement de laquelle l'idée que les lexiques de langage naturel sont bien plus que des collections d'objets indépendants, mais possèdent un mode d'organisation extrêmement riche que nous cherchons à capturer à travers le concept de **paradigme**. Dans cette nouvelle optique, les chaînes graphiques hors-lexiques sont prononcées en deux temps. Une première étape consiste à leur appliquer des réécritures hors-contexte judicieusement choisies, de manière à reconstituer une ou des lexies existantes (les lexies analogues). Ces réécritures, qui correspondent exemplairement à des dérivations morphologiques, sont extraites automatiquement du lexique d'exemples. Le second temps de la procédure consiste à appliquer en retour aux transcriptions des analogues des transformations non-contextuelles dans le domaine phonétique, transformations qui ont préalablement été couplées aux transformations graphiques utilisées dans l'étape de recherche des analogues. Nous montrons les bénéfices d'un tel changement d'optique, aussi bien en terme de modélisation des phénomènes analogiques, de pertinence linguistique, et de qualité des transcriptions produites. Des expériences complémentaires, dans lesquelles une telle démarche est appliquée avec succès à l'apprentissage de la conjugaison des verbes anglais, sont également présentées.

